Applied Data Science Methods in Epitranscriptomic Bioinformatics

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy (or other degree as appropriate) by *Zhen Wei*

29th September 2019

Table of Contents

gcepc: **R package to conduct GC content bias aware peak calling and quantification in meRIP-Seq**

Abstract:

Chemical modifications on messenger RNA have been recently revealed by biological researchers to function as an essential layer of gene expression regulation. Molecular biologists from different laboratories have conducted more than 200 sets of high throughput sequencing experiments trying to capture the types and locations of messenger RNA modifications across multiple cell types and species. However, until this date, the field still lacks a bioinformatics pipeline to quantify and analyze the epitranscriptomic HTS data generated from different laboratories consistently. The thesis aims to provide an overview of questions and challenges arisen in the field of mRNA modification computational analysis. Subsequently, we will present a set of practical computational strategies for data explorations, genomic data mining, modification level quantifications, and technical artifact corrections from a data science perspective. The first chapter of the thesis provides an in-depth data exploration and visualization of m5C mRNA modification from bisulfite sequencing data. In the second chapter, we document the database construction and data consistency exploration for the transcriptomic targets of the mRNA modification related protein regulators. Besides, the second chapter presents a methodological framework for the computational representation of the domain knowledge related to the transcriptomic topology of epitranscriptomic modification. The final section of the thesis discusses the dominant technical biases existed in MeRIP-Seq, the most widely applied type of HTS data in epitranscriptomics, and it follows with a practical computational pipeline to overcome the technical error.

Chapter 1

Topological characterization of human and mouse m^5C epitranscriptome revealed by bisulfite sequencing

1.1 Outline

Background: Compared with the well-studied 5-methylcytosine (m^5C) in DNA, the role and topology of epitranscriptome m^5C remain insufficiently characterized.

Results: Through analyzing transcriptome-wide m^5C distribution in human and mouse, we show that the m^5C modification is significantly enriched at 5' untranslated regions (5'UTRs) of mRNA in human and mouse. With a comparative analysis of the mRNA and DNA methylome, we demonstrate that, like DNA methylation, transcriptome m^5C methylation exhibits a strong clustering effect. Surprisingly, an inverse correlation between mRNA and DNA m^5C methylation is observed at CpG sites. Further analysis reveals that RNA m^5C methylation level is positively correlated with both RNA expression and RNA half-life. We also observed that the methylation level of mitochondrial RNAs is significantly higher than RNAs transcribed from the nuclear genome.

Conclusions: This study provides an in-depth topological characterization of transcriptome-wide m^5C modification by associating RNA m^5C methylation patterns with transcriptional expression, DNA methylations, RNA stabilities, and mitochondrial genome.

Keywords: RNA 5-methylcytosine (m^5C), mRNA methylation, RNA bisulfite sequencing, RNA epigenetics

Abbreviation

m^6A: N6-methyladenosine

hm^5C: hydroxymethylcytosine

Ψ: pseudouridine

m^1A: N1-methyladenosine

m^5C: 5-methylcytosine

BS-Seq: bisulfite sequencing

MDA-MB-468: MDA468

MEF: mouse embryo fibroblast

MEF-Dnmt2$^-$: mouse embryo fibroblast with Dnmt2 knock down

WGBS: whole genome bisulfite sequencing

RRBS: reduced representation bisulfite sequencing

mRatio: methylation ratio

OR: odds ratio

DMS: differential methylation site

lncRNA: long noncoding RNA

sncRNA: small noncoding RNA

1.2 Introduction

DNA methylation is a well-established and extensively studied epigenetic phenomenon [1-4]. In contrast, mRNA methylation is still relatively an uncharted territory [5]. Although the presence of the chemical modifications to tRNA has been established in the 1970s [6-8], little is known about the epigenetic modifications to mRNA and other non-coding RNAs. Even less was known about their abundance, role, and mode of regulation until recently when several studies showed that N6-methyladenosine (m^6A) is the most abundant messenger RNA (mRNA) modification in eukaryotes [9], and suggested to regulate a number of biological processes including translation efficiency [10], circadian clock [11], microRNA processing [12], RNA-protein interaction [13], RNA stability [14], heat shock response [15] and differentiation [16].

Compared to m^6A, even little is known about the abundance and role of transcriptome 5-methylcytosine (m^5C) modification. Existing studies of m^5C in cellular RNA have been largely confined to rRNA and tRNA [17]. For example, RNA m^5C modification in plant

rRNA and tRNA is reported to be conserved [18], and is shown to affect stability of synthetic RNA [19, 20]. In mammalian system, cytosine-5 methylation in tRNA has been shown to regulate Mg^{2+} binding, anticodon stem loop conformation and secondary structure stabilization [21, 22]. In addition, m^5C in tRNAs is reported to regulate protein translation in stress response, tissue differentiation, and neuro-development disorders [23-29]. In rRNA, m^5C is shown to regulate translation process [30]. A recent study also showed that, hm^5C, the intermediate of RNA m^5C demethylation, is enriched in poly-A tailed RNA and the coding sequences of the mRNA transcript, and it is associated with brain development and the active transcription of mRNA [11].

Recent advancement of RNA bisulfite sequencing (BS-Seq) technique [31-34] has enabled the transcriptome-wide m^5C profiling at single base resolution and confirmed its widespread existence in the human transcriptome [34, 35]. Intriguing differences with respect to the degree of transcriptome m^5C methylation, functional classification and position bias were reported with this technique [36], and it was recently shown that transcriptome m^5C promotes mRNA export through methyltransferase NSUN2 and reader ALYREF [37].

It is observed that m^5C modification may accounts for 20% of the total internal methylations on ploy(A) RNA in BHK21 cell line [38, 39]. However, it is not clear that whether the transcriptome m^5C modification is differentially enriched in different cell types, and the topological relationship between RNA methylation and DNA methylation under the same cell lines has not been investigated.

In this study, using BS-Seq approach, we identified transcriptome-wide mRNA m^5C methylome in mouse and human cells. Our results revealed that transcriptome m^5C is enriched and conserved at the 5'UTRs of target transcripts in both human and mouse cells. Interestingly, under all the examined cell lines, we observed a negative correlation

of the methylation patterns between RNA m^5C methylation and DNA m^5C under CpG context, and the RNA m^5C methylations are enriched on mitochondria transcriptome.

1.3 Material and Methods

Sample preparation and RNA bisulfite sequencing

MCF10A normal mammary epithelial cells and MDA-MB-468 breast cancer cells were obtained from ATCC. MCF10A cells were cultured and maintained in DMEM/F12 (Life Technologies, USA) supplemented with 5% Horse serum, EGF (20 ng/ml), Hydrocortisone (0.5 μg/ml), Insulin (10 μg/ml) and Anti-Anti (Life Technologies, USA). Likewise, MDA-MB-468 cells were cultured in RPMI (Life Technologies, USA) supplemented with 10% FBS and Anti-Anti (Life Technologies, USA). For BS-Seq total RNA was isolated from MCF10A and MDA-MB-468 cells and was enriched for polyA+ RNA using poly-A selection kits. The purified RNA is subjected to sodium bisulfite treatment at 60 degrees for 8 hours. The bisulfite-treated RNA was then reverse transcribed and subjected to deep-sequencing using Illumina RNA-Seq protocol. The data has been deposited under Gene Expression Ominous (GEO) with Accession Number GSE84230. To replenish the transcriptome BS-Seq data of aforementioned human samples (MCF10A and MDA-MB-468), additional datasets are obtained from public resources, including DNA BS-Seq data from MCF10A (GEO GSM659628) [40], transcriptome m^5C methylation data from mouse embryo stem cell (ESC) and mouse whole brain profiled by RNA BS-Seq (GEO GSE83432) [41], and mouse ESC DNA methylation data (GSM1873374) [7, 42].

Quality control and alignment of BS-Seq data

The FASTQ files from BS-Seq samples are trimmed with Trim Galore [43], it removes low quality 3’ ends with phred score threshold of 20, and it can remove potential adaptor contamination. Then, the reads are aligned to the reference genomes of mouse and human (mm10 and hg19) with MeRanGs in MeRanTK [44]. The methylation is called using MeRanCall, regions of the 5’ends and 3’ends of the reads are ignored based on the threshold cut-off suggested by the M-bias plot generated by meRanGs. The minimum

reads coverage for the methylation report was set at 10, and the minimum read base quality (phred score) for methylation call is filtered at 30. The maximum reads duplication level is set at 10 to prevent the PCR artefacts; the minimum non-conversion rate to report is set at 0 to include the non-methylated sites as background control for further analysis.

For DNA bisulfite samples, the trimmed reads are aligned using Bismark under alignment setting --score_min L,0,-0.6. The SAM files are filtered by Samtools using -F 1540 and -q 30 to remove reads that are duplicated and whose Quality scores are lower than 30. The methylation status of genome-wide cytosine sites is reported from the filtered SAM files with Bismark methylation extractor with argument --cytosine report. Also, the conversion rate biased ends are also ignored during methylation call based on the M-bias plots. The minimum read coverage was filtered at 10 as well.

Filtering false positive m^{5}C sites due to RNA secondary structure

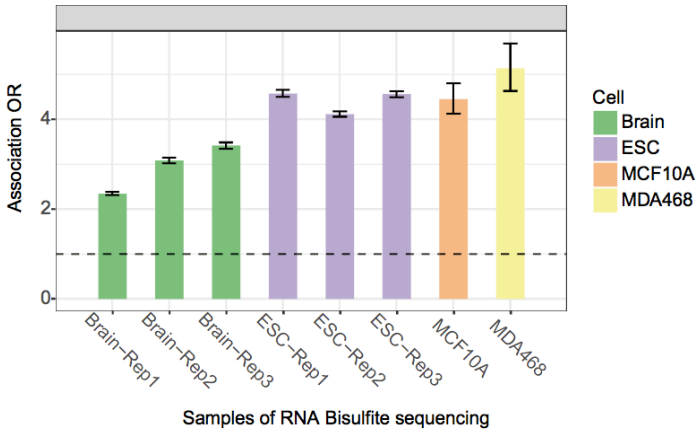


Figure 1. Association between reported methylated sites and double stranded RNA structures before filtering. The reported m5C sites by MeRanTK approach are clearly enriched with double stranded RNA structure before the filtering, and the pattern is consistent in all the 4 samples. The odds ratio (OR) was calculated from Fisher's exact test and the error bar shows the 95% confidence interval of OR estimation.

It is known that, secondary structures on RNAs prohibit bisulfite conversion and thus can result in false positive detection of transcriptome m^5C sites. As shown in **Figure 1**, the detected m^5C sites from meRanTK are enriched with double stranded regions of RNA, which are likely to be false positive errors due to secondary structure. For this reason, an R package rBS2ndStructure was created to facilitate the elimination of the false positive methylation calls due to RNA secondary structures. Specifically, RNA secondary structure is predicted with RNAfold from the Vienna RNA package [45] as it was performed by Amort et.al [41]. The transcriptome wide full length transcripts are extracted from UCSC gene annotation for both mm10 and hg19. Then, the double stranded structures are predicted with the MEA method under alpha = 0.1. The folding temperature is set at 70 degree, and the maximum pairing distance is set at 150bp. For the mitochondria chromosome and the transcripts longer than 8000bp, the structures are predicted using sliding windows of 2000bp and step size of 1000pb. For both the RNA and the DNA methylation reports, the methylation sites overlapped with the predicted regions of secondary structures are filtered. Due to the lack of computational resources to predict structures on large intronic sequences, the cytosine sites that do not locate on the exons of known transcripts or the mitochondria chromosome are filtered. The resulting methylation reports are then analysed under R environment using primarily GenomicFeatures [46], Guitar [47] and ggplot2 [48] packages.

The rBS2ndStructure package is publically available at Github (https://github.com/ZW-xjtlu/rBS2ndStructure) with precomputed RNA secondary structures of genome assembly mm10 and hg19 for convenient processing of RNA BS-Seq result.

Quantitative analysis of methylation status

The methylation ratio (mRatio) of a specific Cytosine site is calculated by:

$$\text{methylation ratio} = \frac{\#\text{ of unconverted Cs}}{\#\text{ of unconverted Cs} + \#\text{ of converted Cs}} \quad (1)$$

where “# of unconverted Cs” and “# of converted Cs” indicates the count of methylated (un-converted) Cs and unmodified Cs (converted Cs) at a specific Cytosine site, respectively. The methylation rate is conceptually similar to the well adapted concept of “beta value” in DNA methylation analysis [49], which indicates the percentage of methylated Cs among all Cs. Also, it is not difficult to show that

$$\begin{aligned} \text{methylation ratio} &= \frac{\#\text{ of unconverted C}}{\#\text{ of unconverted C} + \#\text{ of converted C}} \\ &= 1 - \frac{\#\text{ of converted C}}{\#\text{ of unconverted C} + \#\text{ of converted C}} \\ &= 1 - \text{convertion rate} \end{aligned} \tag{2}$$

where the conversion rate has been previously defined in [35] and a smaller value suggests higher percentage of RNA m^5C methylation.

To differentiate a set of statistically significantly methylated Cytosine sites against potential technical randomness due to incomplete bisulfite conversion, the p values for the methylation state of both the DNA and RNA methylation are calculated by fisher exact test which against the background conversion odds after the filtering of the sites mapped to introns and secondary structures. The adjusted pvalues (FDR) are then adjusted by Benjamin & Hochberg method. The positive methylation states were decided when FDR < .05.

For the mouse samples containing 3 biological replicates, the methylated sites are judged as FDR <.05 among all 3 replicates. For other insignificant methylated sites to be kept in the analysis, the sites should be reproduced 3 times with coverage > 10. The converted reads and non-converted reads are added on each site when combining the biological replicates.

The background bisulfite non-conversion rate is 2.75%, 2.74%, 1.18% and 0.81% for MCF10A, MDA468, mouse ESC and mouse brain$^-$ samples, respectively (taking average for samples with more than one biological replicates). The difference among non-

conversion rates might be due to the biological difference of cell-lines, batch variation and different BS-Seq protocols.

Differential methylation analysis

The odds ratio (OR) or methylation fold change from differential analysis is defined as:

$$\underset{\text{from differential methylation}}{\text{odds ratio}} = \frac{\left(\frac{\#\text{ of unconverted Cs under cond\_1}}{\#\text{ of converted Cs under cond\_1}}\right)}{\left(\frac{\#\text{ of unconverted Cs under cond\_2}}{\#\text{ of converted Cs under cond\_2}}\right)}. \quad (3)$$

Odds ratio (or methylation fold change) indicates whether the methylation is enriched under one condition compared with another condition. A value greater than 1 suggests increased methylation level, where as a value less than 1 suggests decreased methylation level. The statistical significance of the odds ratio is evaluated by QNB method, which tests the homogeneity of association between methylated and unmodified molecules under two experimental conditions with the within-group variability assessed through 4 cross-linked negative binomial distributions [50].

Similar to the odds ratio from differential methylation analysis, the enrichment odds ratio of m^5C sites within a specific region can be defined as:

$$\text{enrichment odds ratio} = \frac{\left(\frac{\#\text{ of } m^5C \text{ sites within a region}}{\#\text{ of C sites within a region}}\right)}{\left(\frac{\text{total \# of } m^5C \text{ sites}}{\text{total \# of C sites}}\right)} \quad (4)$$

A value greater than 1 suggests that methylation sites are enriched within the tested region. And the statistical significance of enrichment can be evaluated by Fisher's exact test. Please note that, in this analysis, we used the total number cytosine sites reported from meRanTK rather than total number of all 4 types of nucleotides.

Assessing the distribution of m^5C sites on mRNA

The distribution pattern of m^5C sites on mRNA is assessed with the Guitar R/Bioconductor [47]. Compared with other software tools and methods, Guitar package provides an improved resolution by relying on only the mRNA transcripts that have sufficient long (more than 100bp) 5'UTR, CDS and 3'UTR simultaneously. For instance, transcripts without annotated 5'UTRs will be excluded from the analysis. Additionally, Guitar doesn't rely on only the primary transcript (often defined as the longest transcript among all isoforms in practice) when solving ambiguous association between a m^5C site and the isoform transcripts of a gene; instead, all ambiguous associations are considered with the weight of association evenly divided. For example, if a single m^5C site locates on 3'UTR of a transcript and CDS of another isoform transcript of that gene, it is counted as half of the m^5C site is located on 3'UTR and the other half located on 5'UTR. In this way, the isoform information is largely retained. To our knowledge, Guitar package should provide the most accurate assessment of transcriptomic distribution pattern.

Differential expression analysis

Differential expression analysis was performed with DESeq2 package [51] and the aligned RNA BS-Seq data.

Cell culture and viral infection

Jurkat T Lymphocytes were maintained in RPMI 1640 medium (Hyclone) supplemented with 5% (v/v) FBS (Gibco) and 100U/ml penicillin/streptomycin (Hyclone). For infection, Jurkat cells were infected with known amounts (3 X 108 genome copies per 2 X 105 cells) of SRV for 18 hours at 37℃, followed by three times wash with PBS (Hyclone). Infected cells were incubated in completed culture medium for the indicated time. Successful infection was identified as the appearance of cytopathic effects in infected cells at 8 to 10 days postinfection.

Reverse transcription and realtime PCR

SRV genome in culture medium was extracted by viral RNA extraction kit (TIANGEN) and reverse transcripted into cDNA by reverse transcriptase PCR kit (TaKaRa). Cellular genome was extracted by TIANamp Genomic DNA Kit (TaKaRa). Realtime PCR was performed in 7500 Fast Real-Time PCR System (Applied Biosystems) by using Premix Ex Taq (Probe qPCR) kit (TaKaRa). SRV genome positive control, primers and probe; as well as GAPDH primers and probe were kindly provided by VRL China Ltd [52].

Immunofluorescence assay

Cells were seeded on poly-L-lysine (Sigma) coated slides, fixed with 4% paraformaldehyde for 15 minutes, followed by permeabilized with pre-cold pure methanol for 20min at -20℃ and blocked with 5% BSA for 1 hour. Cells were then stained with the serum from SRV infected monkey (1:25 diluted in blocking buffer) overnight and visualized with DyLightTM 488-Labeled anti-Human antibody (KPL). Cells were counterstained with Hoechst (Life Technologies) for 10 minutes and mounted on microscopy slides. Samples were imaged with a ZEISS LSM 880 Confocal Laser Scanning Microscope.

1.4 Results

Overview of mRNA m^{5}C methylome revealed by BS-Seq

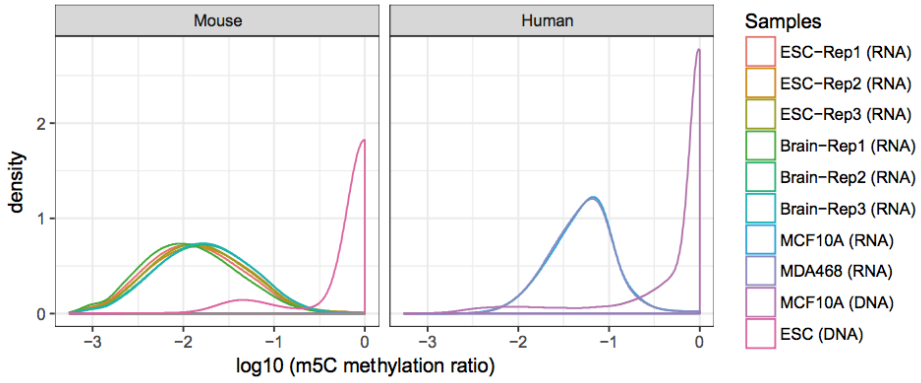
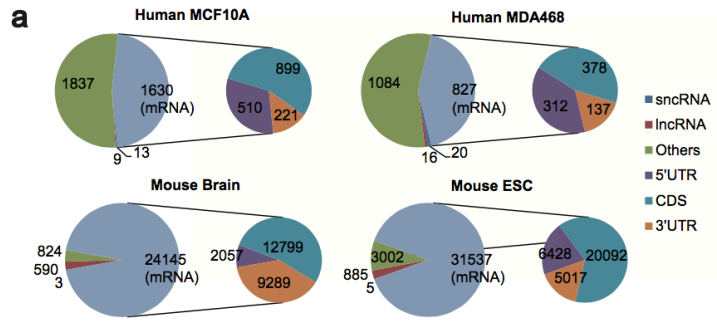


Figure 2. Distribution of RNA and DNA methylation ratio. The diagram shows the distribution of RNA and DNA methylation ratio under 4 different conditions. In general,

DNA methylation ratio is much higher than RNA. The distribution patterns of biological replicates are relatively close.

After successful processing of the RNA BS-Seq datasets, a total of 3440 (0.40%), 1915 (0.29%), 35246 (0.757%) and 25301 (0.50%) RNA cytosine sites were identified as m^5C methylation (FDR < 0.05) sites in MCF10A, MDA-MB-468, mouse embryonic stem cell (ESC) and mouse whole brain, respectively. The overall transcriptome m^5C methylation level was much lower than DNA m^5C methylation level (**Figure 2**). Importantly, we found that m^5C was widespread in different RNA families, where more than 50% of them were located on mRNA (**Figure 3a**). In MCF10A cells, 7131 protein coding genes had sites reported after the filtering, of which 225 (3.15%) mRNAs contained m^5C sites. In MDA-MB-468 cells, 6320 protein coding genes had reads aligned, of which 128 (2.06%) contained m^5C sites. In ESC and brain- samples, the methylation status were available for 11325 and 13108 protein coding genes, of which 3579 (31.6%) and 3065 (23.4%) contained m^5C sites. The difference in number of m^5C sites between different conditions is mostly due to different sequencing depth.



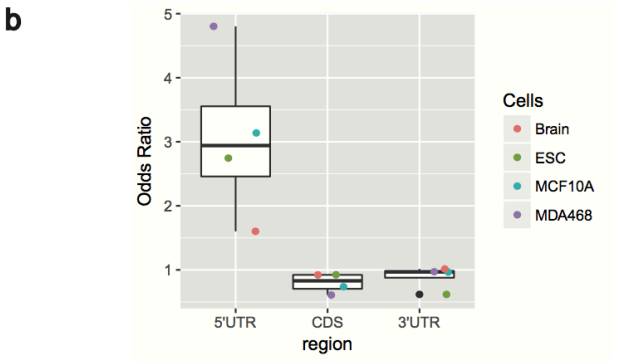


Figure 3. Distribution of transcriptome m^5C modification sites in human and mouse. **(a)** The pie chart shows transcriptome wide distribution of m^5C sites in MCF10A, MDA-MB-468, mouse embryonic stem cell (ESC) and whole brain. The majority of the identified m^5C sites are located on mRNAs. **(b)** Graph showing status of m^5C frequency in different regions of mRNA. The result indicates that, detected Cytosine sites are consistently enriched at 5'UTR on mRNA compared with CDS and 3'UTR.

mRNA m^5C is enriched in 5'UTRs of human and mouse

To study the spatial organization of m^5C sites in the transcriptome, we first analysed the relative enrichment (see methods for more details) of m^5C sites on different types of RNA and at different regions (shown in **Figure 3b**) by compensating for the Cytosine sites that do not carry m^5C modification. Our results showed that m^5C sites were consistently and significantly enriched at 5'UTRs in human and mouse with enrichment odds ratio of 3.138, 4.802, 2.744 and 1.601. The similar topology was already reported by the previous studies [35, 36], and our observation further confirm their conclusions. Also we did observe a slight enrichment of m^5C sites in 3'UTR in mouse brain (enrichment odds ratio = 1.013 and 1.19E-02), which is also reported in the study of *Almert.et al [36]*, 3'UTR enrichment was not observed in the other samples (odds ratio = 0.964, 0.971 and 0.617).

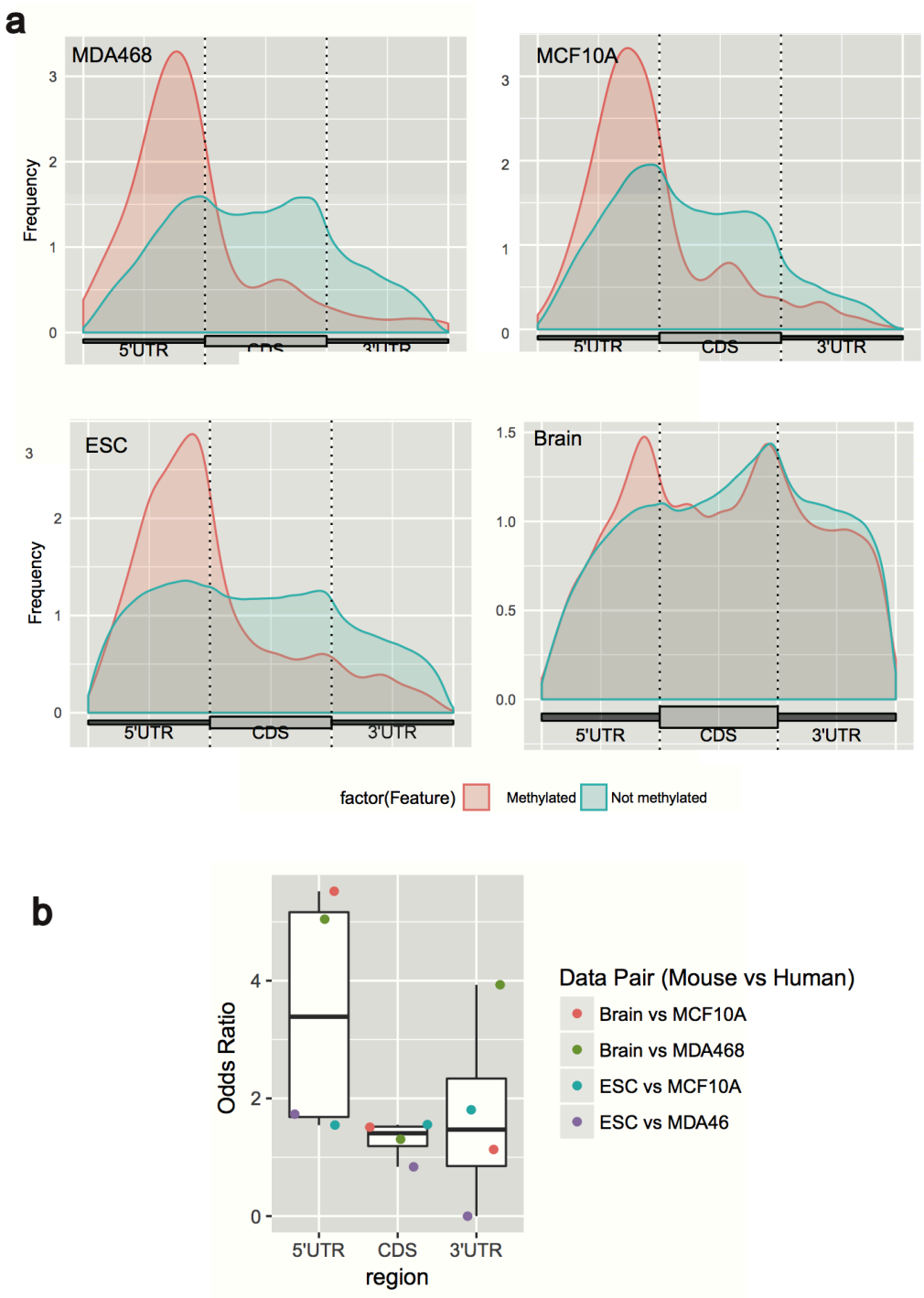


Figure 4. Conservation of m^{5}C in different mRNA regions. (b) Graph showing status of m^{5}C frequency in different regions of the transcripts. We divided all the detected Cytosine sites into 2 groups based on whether it is methylated. The result indicates that, Cytosine sites with significant methylation levels are consistently enriched at 5'UTRs and near start codon in all 4 samples. The m^{5}C sites in mouse brain are peaked at 3'UTR

and CDS compared with other samples. However, the background sites in mouse brain samples are also peaked around 3'UTR and CDS. Therefore, the relative enrichment of the mouse brain sample is still at the 5'UTR, as it is shown in figure 3b. The differences in the distributions of background sites might be attributed to substantial differences in the transcript expression profiles between the mouse brain samples and other samples. **(b)** A correlated methylation pattern is observed on 5'UTRs between different cell lines/tissues in human and mouse. The conserved Cytosine residuals were retrived with liftOver utility (http://genome.ucsc.edu/cgi-bin/hgLiftOver). And the correlation analysis is performed with a Fisher's exact test. It is important to note that, although we failed to observe correlated m5C methylation pattern on CDSs and 3'UTRs of mRNA across all 4 pairings, it is possible that such pattern may emerge on strictly matched cell lines/tissues.

To further substantiate these findings, we plotted the distribution of the methylated and un-methylated Cytosine sites locate on mRNAs with Guitar package [47]. In order to improve the resolution of this analysis and differentiate the distribution of m^5C sites on usually short 5'UTRs, only the mRNAs with a 5'UTR longer than 100 bp are used. As shown in **Figure 4a**, the methylated Cytosine sites were consistently enriched at 5'UTRs across all 4 samples when compared to un-methylated groups. Interestingly, this trend is also supported by the Cytosine methylation sites reported by Squires *et al.* [35], and there is no significant enrichment of m^5C sites observed in 3'UTR when all cytosine methylated were used as background (**Figure 5**).

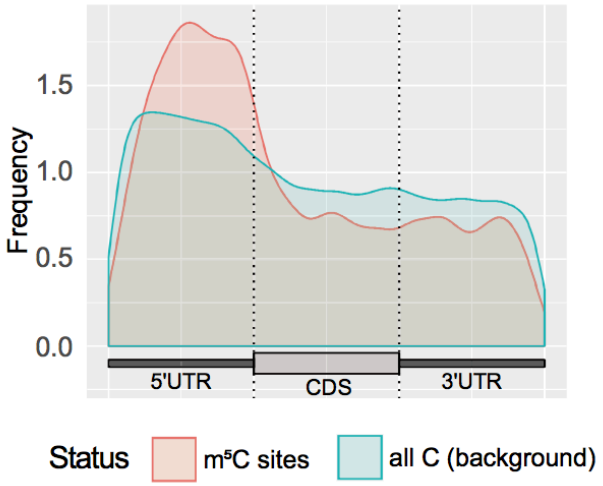


Figure 5. mRNA m^5C sites are enriched at 5'UTRs in HeLa cells. The 10275 RNA methylation sites from Hela cell line are extracted from a previous study (1). Compared with the background (all the transcriptomic Cytosine sites), the reported mRNA m^5C methylation sites are more enriched on 5'UTR. Prominent enrichment is not observed on 3'UTR. This trend is consistent with the other 4 samples tested (See Figure 1 of the main text). Figure is plotted using default setting of Guitar package (2) with human hg19 genome assembly and UCSC gene annotation.

We further compared the methylation status of the conserved locus in human and mouse between different cell lines/tissues. After pairing the cytosine sites of human and mouse by LiftOver, the correlation of m^5C sites between species is quantified using the odds ratio association of the significant modification sites against background sites between 2 samples under a given transcript region. The odds ratio value larger than 1 indicates a positive association on m^5C positive sites between species, since it suggests a tendency for evolutionary conservation. It is observed that, although the cell types/tissues we used are not strictly matched, a consistently positively correlated methylation pattern is observed on 5'UTR region (**Figure 4b**). However, unlike 5'UTR, correlated pattern of m^5C sites were not consistently observed in CDSs or 3'UTRs in our study, the observed heterogeneity of the m^5C methylome in different transcript regions suggests that the m^5C mapped to 5'UTR of the transcripts are more likely to be

functional important. High degrees of associations are also observed in Brain vs MDA468 and ESC vs MCF10A under 3'UTR. This pattern suggests that some of the m^5C sites on 3'UTR are evolutionary conserved, although the conserved sites may not be consistently expressed under all cell line conditions.

m^5C site exists under different nucleotide contexts

Because RNA methyltransferase Dnmt2 shares strong sequence homology to all DNA DNMT methyltransferases [53], we reason that exploring the relationship between transcriptome and DNA m^5C methylation profiles may unravel interesting interplay between the two kinds of reversible chemical modifications. In mammalian cells, DNA methylation occurs mainly at CG dinucleotides (including ACG, CCG, TCG and GCG, See **Figure 6b**). To study whether, like DNA methylation, transcriptome m^5C methylation also occurs at the similar position, we analysed methylated cytosine in the transcriptome. For this purpose, we examined the all possible C-centred trinucleotide combinations. Unlike DNA, transcriptome m^5C occurs at all C-centred trinucleotides (**Figure 6a**), and was observed to be specifically enriched at GCA, ACG, CCG, GCG, CCC, and GCC. These results were found to be consistent within the same species (Pearson correlation = 0.96 and 0.92, **Figure 6c**) and between different species (Pearson Correlation = 0.72, 0.75, 0.45 and 0.48, **Figure 6c**).

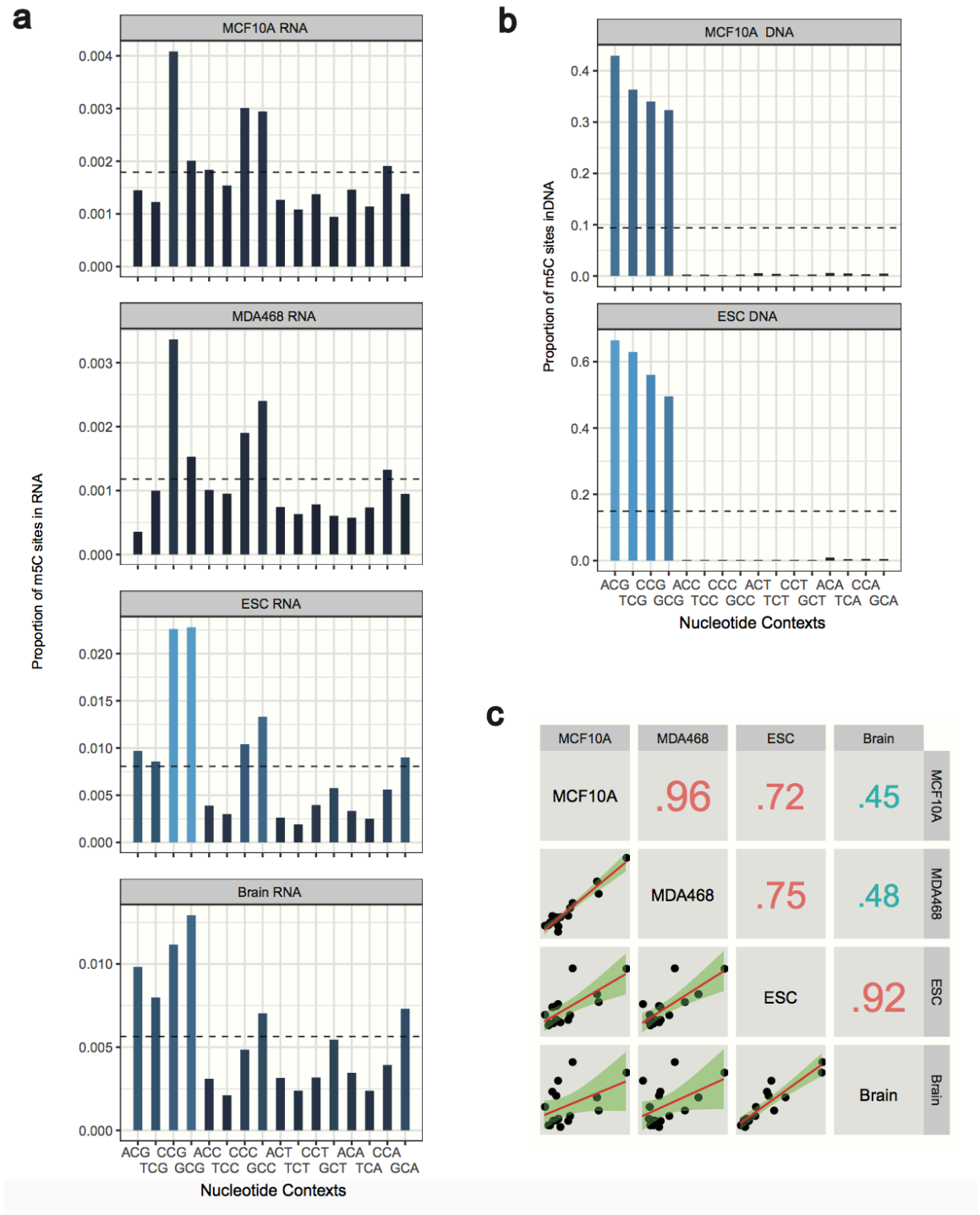


Figure 6. Comparative distributions of mRNA and DNA m^{5}C methylation. (a) Bar graph shows the proportion of mRNA m^{5}C sites under different combination of C-centered trinucleotides in mouse and human cells. The dotted line shows the average percentage of methylation under all trinucleotide contexts within the entire transcriptome.

We observed that RNA m^5C occurs under all trinucleotide contexts and showed slightly enriched in sequences containing CCG, GCG, GCC, GCU and GCA. **(b)** Bar graph showing proportion of DNA m^5C sites in mouse and human cells. DNA cytosine sites were enriched exclusively in sequences containing CG dinucleotides (ACG, CCG, CCG and TCG). **(c)** The coefficient of correlation between RNA methylation and trinucleotide sequences was found to be consistent between samples from the same species (Pearson correlation = 0.96 for human and 0.92 for mouse) and also between human and mouse cells (Pearson correlation = 0.72, 0.75, 0.45 and 0.48).

Negative correlation in methylation level is observed between mRNA and the corresponding exonic region of DNA

We next examined whether there exists any correlation between m^5C methylated/non-methylated (m^5C methylation ratio) in transcriptome and the corresponding DNA exonic regions at each C-centred trinucleotide sites. Because DNA methylation occurs mainly at CG dinucleotides, as expected we observed no strong correlation at non-CG trinucleotides. However, we observed significant negative correlation in methylation ratios between RNA and DNA at all four CG-containing trinucleotides. As a higher percentage of m^5C in mRNA is detected, the corresponding DNA exonic CG dinucleotide was less likely to be methylated (**Figure 7a**). Next, we grouped m^5C methylated at all CG sites according to their methylation ratio (methylated and unmethylated) and investigated their distributions in mRNA and the corresponding exonic regions of DNA. Consistent with our previous finding, we observed significant negative correlation in both human and mouse cells. In particular, 5'UTR in mRNA showed high methylation ratio, whereas corresponding DNA region showed significantly low methylation ratio (**Figure 7b**).

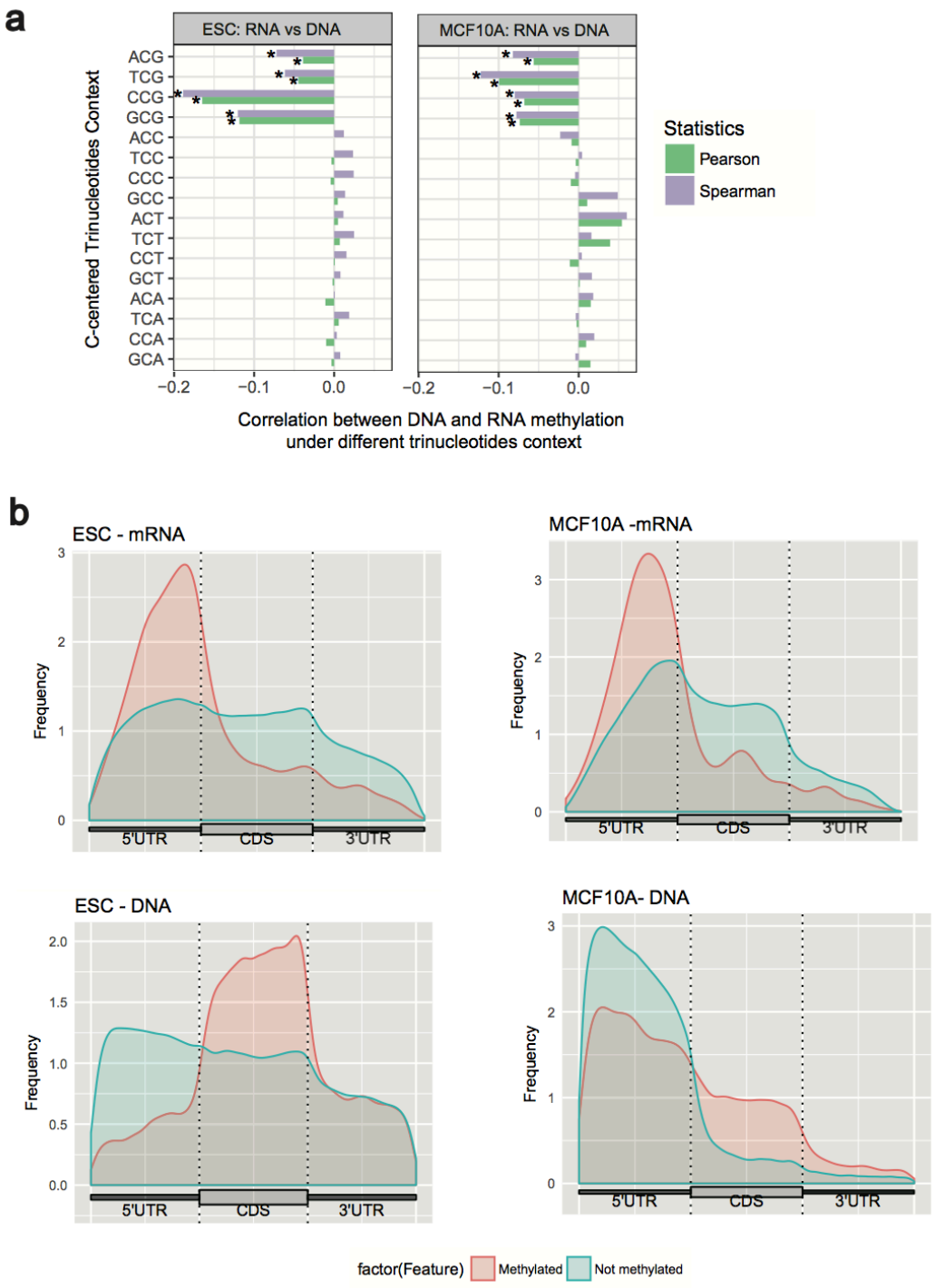


Figure 7. The methylation ratio of corresponding m^{5}C DNA and mRNA CpG islands shows negative correlation. (a) Negative correlation is observed between DNA and mRNA methylation ratio consistently under all four CG containing trinucleotide (ACG, TCG, CCG and GCG) in both human and mouse, i.e., if a specific CG dinucleotide in

DNA is methylated, the corresponding dinucleotide in mRNA is significantly less likely to be methylated. "*" labeled the top the 4 nucleotide contexts under which strongest correlation between DNA and RNA methylation level exists. **(b)** Comparative distributions of m^5C methylated CG sites in DNA and RNA show an enrichment of sites with high methylation ratio in mRNA 5'UTR as opposed to an enrichment of low methylation ratio sites in DNA 5'UTR. The pattern is consistent in both human MCF10A cell line and mouse embryo stem cells.

Transcriptome m^5C sites exhibit a clustering effect

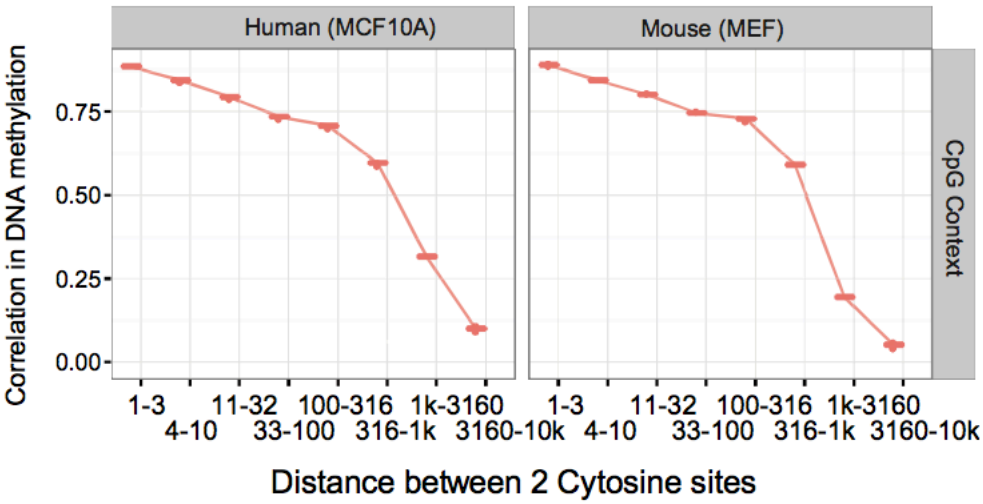


Figure 8. Strong clustering effect exists in DNA methylation under CpG context. Figure shows the correlation in methylation level between two cytosine sites of a certain distance. Strong correlation is observed between cytosine sites with a smaller distance.

In DNA methylation, it has been shown that correlation of methylation rates between two CpG sites is related to the distance (see **Figure 8**), and the clustering effect can be as high as 0.7 for probes within 200bp [54]. To address whether the mRNA m^5C methylation also exhibits clustering effect, we examined the proportion of m^5C sites that are within 10bp distance of another m^5C sites, and compared this proportion with that from 1000 times of random permutation. Our analysis revealed that m^5C showed

obvious clustering effect in both mRNA and DNA (**Figure 9a** and **Figure 9b**). In ESC cell line, more than 76.7% of the mRNA methylation sites had at least one methylation site mapped within 10nt flanked region, compared with 7.7% of such event by random permutation of methylation states on insignificant methylation sites of the methylated genes. In mouse ESC and Brain cells, more than 43.02% and 30.06% of mRNA m^5C methylation sites existed within the m^5C-p-m^5C dimmers, compared with expected rate of 1.02% and 0.77% of such dimmers by the random permutation.

To further elucidate the clustering effect, we calculated the correlation of methylation ratio between two cytosine sites with a specific distance. To our surprise, mRNA methylation exhibited a stronger clustering effect compared with DNA (**Figure 9c**). In addition, the correlation of methylation ratio was consistently stronger within 1-3nt distance as revealed of higher correlation of methylation ratio (0.76 in MCF10A and 0.79in ESC). These results indicated that most CpC dimer are co-methylated, the correlation of methylation ratio can be as high as 0.58 in MCF10A and 0.47 in ESC for Cytosine sites with a distance of 4-10nt. Though the overall clustering effect of DNA methylation was not as strong as mRNA methylation, when only CpG dinucleotide was considered, DNA methylation exhibited a stronger clustering effect than mRNA methylation (see **Figure 8**).

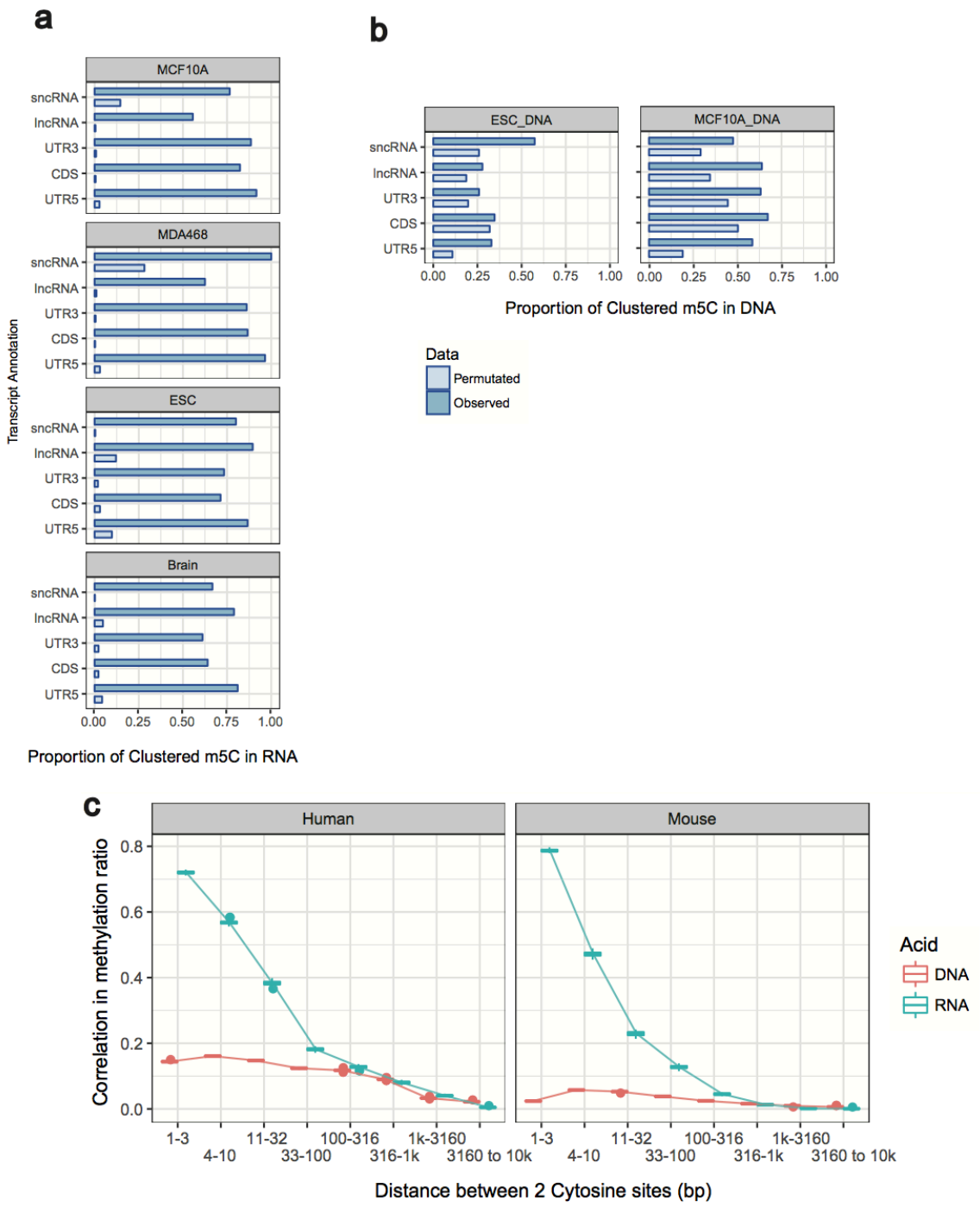


Figure 9. RNA m^{5}C modification exhibits clustering effect. (a) Bar graph shows the proportion of clustered m^{5}C sites within 10nt flanked regions. To evaluate the statistical significance, we generated 1000 permutated results as a comparison with the bars indicating a 99% confidence interval. Using these criteria, m^{5}C methylation showed a strong clustering effect consistently on different RNA families and on different regions of mRNA in human and mouse. Around 50% of the m^{5}C sites were clustered with each

other within a 10bp region. **(b)** DNA methylation also exhibited a clustering effect. However, the pattern is not that strong when all nucleotide contexts are considered **(c)** Line graph showing correlation between RNA/DNA m^5C methylation and distance between cytosine sites. RNA m^5C methylation showed strong correlation with Cytosine sites that are immediately close to each other. The clustering effect of DNA methylation is strong when only CpG context is considered (**Figure 8**).

Transcriptome m^5C is strongly enriched in mitochondrial transcripts

To further establish a physiological relevance of m^5C distribution, we examined the methylation level of RNAs encoded in different chromosomes. Surprisingly, m^5C modification was strongly enriched in RNAs transcribed specifically from mitochondrial DNA in normal and breast cancer cells as well as in mouse ESC and brain as revealed by enrichment odds ratio of 818.42949, 634.72723, 1028.52065 and 67.28553, respectively. In contrast, the enrichment odds ratios of RNA methylation for transcripts from other chromosomes were found to be roughly the same (**Figure 10a**). The RNA transcripts of all the major genes located on mitochondrial chromosome were significantly methylated (**Figure 10b**). Previously, it was reported that methyltransferase NSUN5 can regulate mitochondrial gene expression [55], and we speculate that RNA m^5C may play a more

vital regulatory role in mitochondria related biological processes.

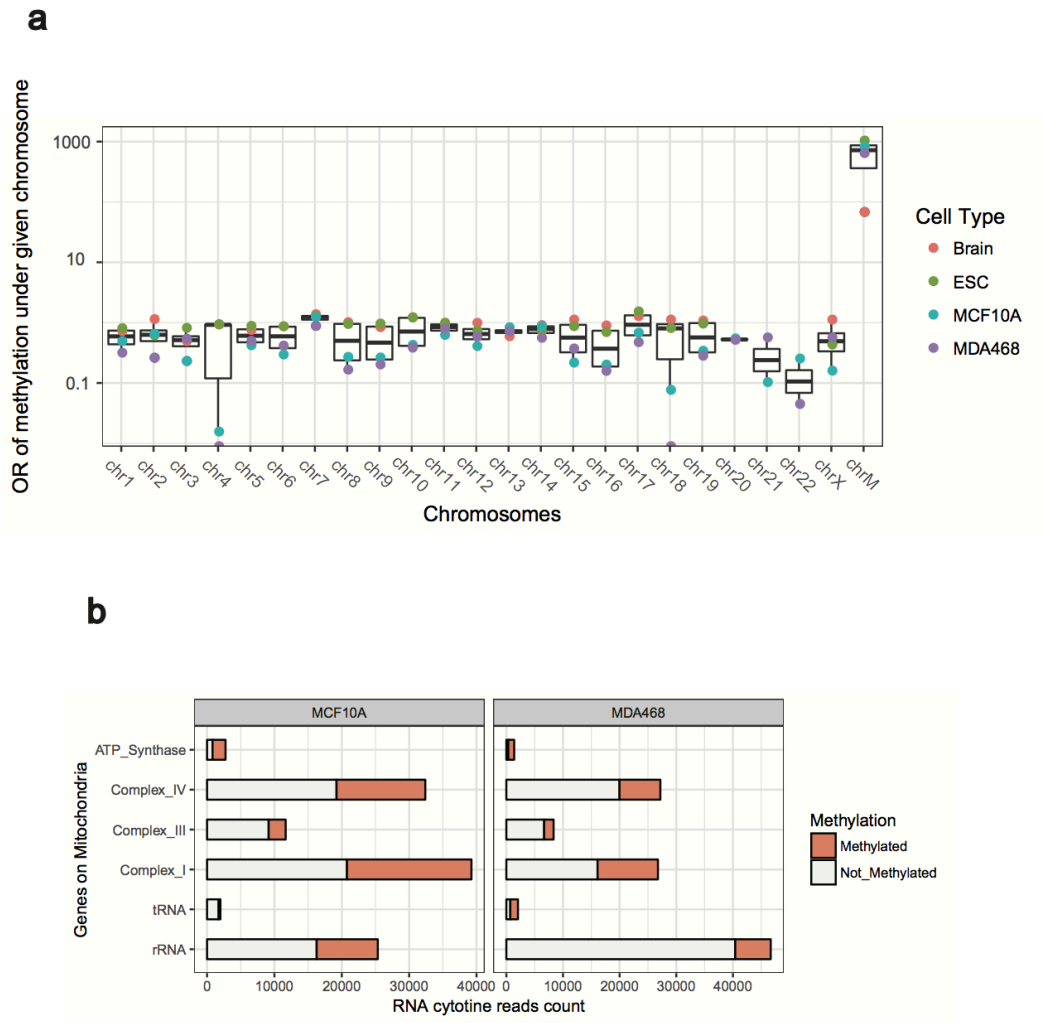


Figure 10. m^{5}C is enriched on mRNAs transcribed from mitochondrial DNA. (a) Bar graph depicting m^{5}C mRNA methylation sites on different chromosomes. RNAs transcribed from mitochondrial DNA (M) showed drastically increased frequency of m^{5}C sites (enrichment odds ratio of 818.42949, 634.72723, 1028.52065 and 67.28553). (b) Bar graph showing the number of methylated cytosine reads stacked with unmodified cytosine reads generated from 6 major classes of mitochondria genes. The RNA transcripts of all the major genes located on mitochondrial chromosome were significantly methylated.

Dys-regulation of RNA methylome in breast cancer

Comparison of normal (MCF10A) and breast cancer (MDA-BM-468) m^5C epitranscriptome identified 162 significant differential methylation sites (DMSs) located on 47 annotated genes at significance level of 0.05. Among the 47 differentially methylated genes, 35 shows hypo-methylation and 12 shows hyper-methylation in cancer cells compared with normal control cell line. The majority differential methylation sites are hypo-methylation (**Figure 11a**), and the m^5C hypo-methylations are mostly located in CDS and 3'UTR region of mRNA but not in 5'UTR region (**Figure 11b**). We then investigated whether different m^5C mRNA methylation levels in normal and breast cancer cells have any functional correlation. We performed functional gene set enrichment analysis on genes containing DMS using DAVID [113] web server and found that many of the 47 differentially methylated genes are related to important biological functions of cancer, e.g., regulation of apoptosis and programmed cell death with RTN4, NME2, CASP14, HSPB1, RPL11 and RPS3 differentially methylated.

Interestingly, like the difference between breast cancer cell line MDA-MB-468 and normal epicelial cell line MCF10A, mechanistic similar mouse stem cells [56] also exhibit dominant hypo-methylation in m^5C epitranscriptome when compared with mouse brain cells with 2513 genes hypo-methylated and 767 genes hyper-methylated (**Figure 11c**). Also similar to previous case, the hypo-methylations are mostly located in CDS and 3'UTR regions of mRNA, but not in 5'UTR region (**Figure 11d**). Using DAVID, we found that hyper-methylated genes in ESC cells are mostly enriched with regulation of cell cycle (FZR1, E2F5, BOP1, TRRAP, CDK4, JUNB, etc), cell death (SIVA1, MCL1, YPEL3, ARF6, UBQLN1, SHF, CIAPIN1, APLP1, GPX1, CASP3, etc.) and mRNA metabolic process (SCAF1, FIP1L1, STRAP, RBM15B, CWC15, XAB2, YBX1, AUH, SF3B2, APLP1, HNRNPL, tec.); the hypo-methylated genes are enriched with functions related to ATP synthesis (ATP6V1F, ATP6V1C1, ATP6V0C, ATP6V1A, ATP6V0E, ATP6V1E1, ATP5C1, etc.), and mitochondrial ribosome (MRPL15, MRPL27, MRPL16, MRPL36, MRPL39, MRPL34, DAP3, etc.). These

results may suggest that the m5C methylations are selectively methylate on transcripts having biological functions related to the cell line conditions.

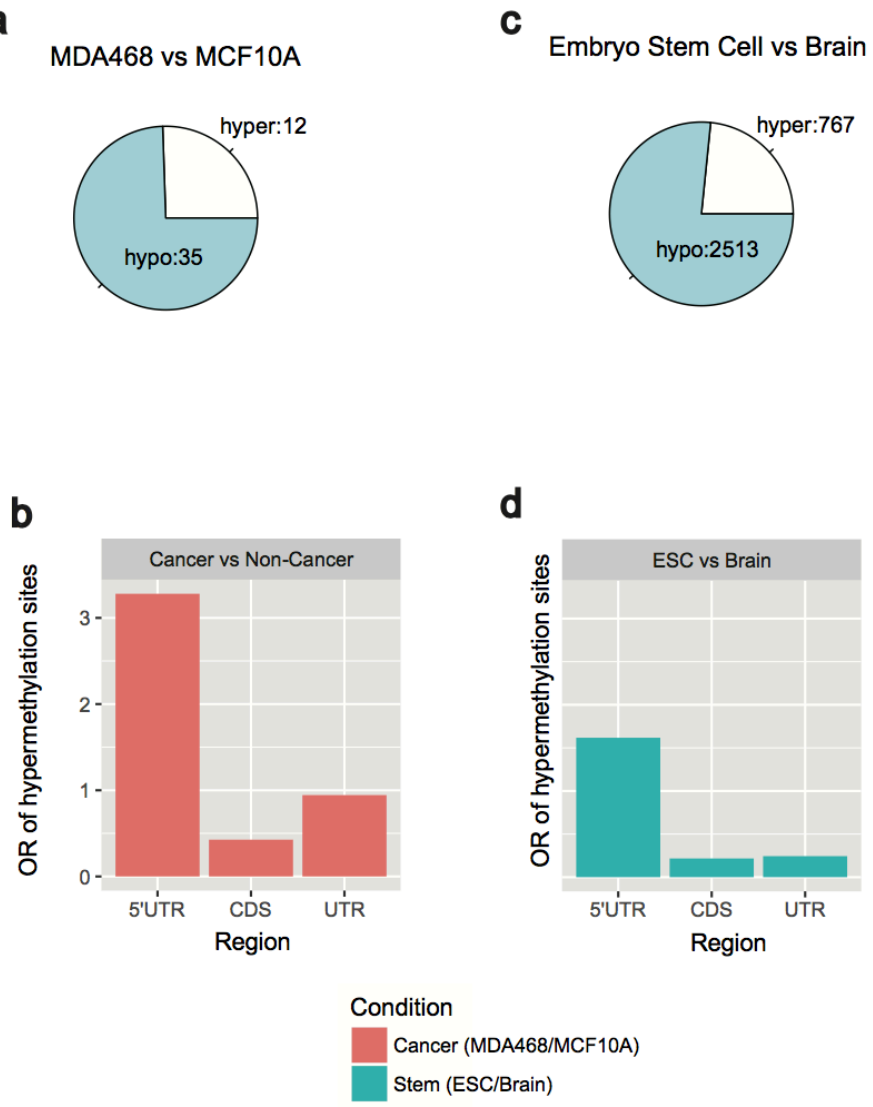


Figure 11. Differential m^{5}C mRNA methylation in different tissues. (a) Pie-diagram showing hypo- and hyper-methylation in MDA468 when compared to MCF10A. A total 47 differential methylated genes were identified between breast cancer (MDA-MB-468) and normal control cell line (MCF10A) with primary hypo-methylation under cancer condition. **(b)** Bar graph showing odds ratio of hyper-methylation sites with respect to all differentially methylated sites on different regions of mRNA. Hyper-methylated sites

were strongly enriched in 5'UTRs. **(c)** Pie-diagram showing hyper-methylation in mouse embryo stem cells when compared to whole brain cells. **(d)** Bar graph showing odds ratio of hyper-methylation sites with respect to all differentially methylated sites on different regions of mRNA in the mouse experiment. Hyper-methylated sites were strongly enriched in 5'UTRs.

Positive correlation between m^5C mRNA methylation and expression changes

In our data, as the gene expression is also estimated from RNA bisulfite sequencing data, a direct comparison of expression and m^5C methylation changes may be problematic due to dependent noise. To eliminate the interference of dependent noise between expression and methylation data, the samples are further divided for different purposes. Specifically, the 3 biological replicates are divided into 2 groups, with the 1 sample used for estimation of expression changes and the other 2 samples for estimation of methylation changes. The expression changes and methylation changes are then compared. This procedure was repeated for 3 times using different grouping combinations.

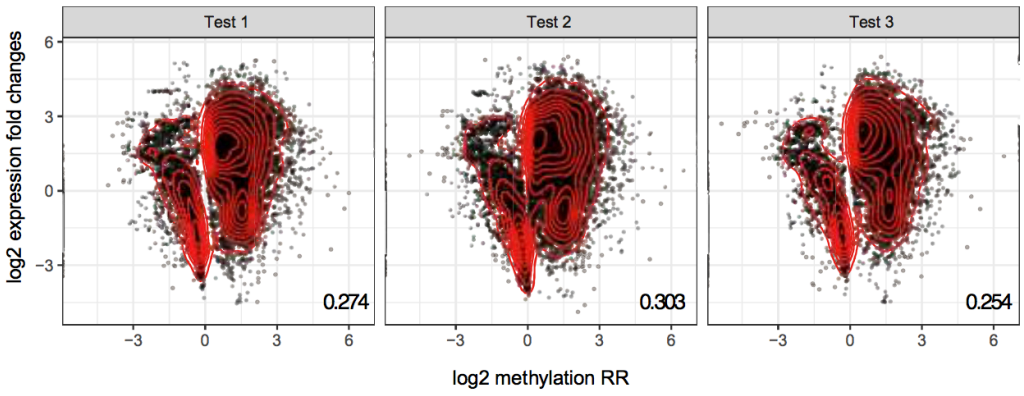


Figure 12. Positive correlations are observed between expression and mRNA m^5C methylation fold change. To eliminate the interference of dependent noise between expression and methylation data, the samples are further divided for different purposes, and the test was repeated for 3 times using different ways of sample grouping. A strong, consistent and significant positive correlation is observed (0.274, 0.303 and 0.254) between log2 expression fold change and log2 methylation fold change when comparing mouse embryo stem cells with brain cells, suggesting that increased methylation level is

likely to be associated with increased expression level, but the underlying mechanism is not yet clear.

A consistent and significantly positive correlation is observed (0.274, 0.303 and 0.254) between log2 expression fold change and log2 methylation fold change when comparing mouse embryo stem cells with brain cells (**Figure 12**), suggesting that increased methylation level is likely to be associated with increased expression level. Although the specific molecular mechanism is not yet clear, the observed positive correlation between RNA m^5C and RNA expression confirmed our previous observed anti-correlation between DNA and RNA m^5C methylation (see **Figure 7**) from a different perspective.

To explain the positive correlation between expression and transcriptome m^5C methylation, we compared the methylation status of all the genes and their half-life, where half-life of mouse genes were obtained from a previous study [57]. The mRNAs are classified into two groups based on whether they have at least one m^5C site or not. To exclude the confounding factor (effective size in methylation site calling), a generalized linear model of binomial family was fitted to half-life with both expression and methylation information. Our result suggests that, there exists a significant positive correlation (pvalue = 2.23e-12) between mRNA half-life and its m^5C methylation status in mouse embryo stem cells and the positive association is also confirmed on mouse whole brain dataset (pvalue = 0.0374). To further exclude the impact of mRNA expression in calling methylation status, we also extracted the genes whose log2 expression levels fall between 7 and 11, and then fit their mRNA half-life with a local regression. As shown in **Figure 13**, compared with the genes of a similar expression level but without an m^5C site, the half-life of the mRNAs that carry m^5C sites is clearly longer and the pattern is consistent in both mouse brain and ESC. This observation may contradict to the global hypo-methylation pattern demonstrated in the comparison between MDA468 and MCF10A, since most of the gene expressions in cancer cells will be upregulated. However, different from the global hypo-methylation, the m^5C on 5'UTR are hyper-

methylated in MDA468, which will suggest the contribution of RNA stability in cancer through m^5C sites under 5'UTR regions.

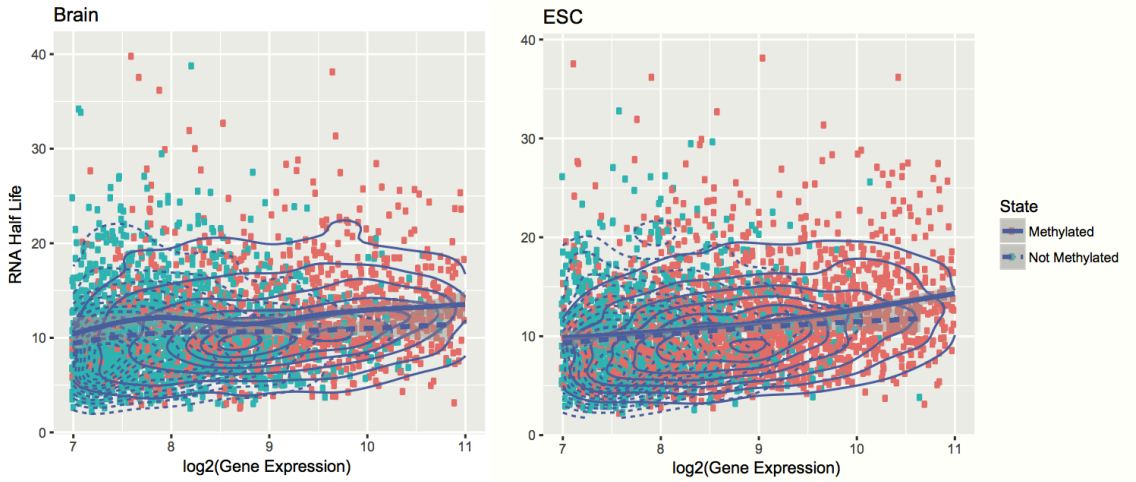


Figure 13. RNA m^5C status is positively correlated with RNA half-life. In the above figure, each red dot represents a gene that carries m5C sites, and each blue dot represents a gene that does not carry an m5C sites. When comparing the methylated and unmethylated genes of similar expression, the genes that carry an m5C site have longer RNA half-life than those do not carry m5C sites. **(a)** Positive correlation between RNA methylation status and RNA half-life is observed in mouse brain (pvalue = 0.0374, generalized linear model of binomial family). **(b)** Positive correlation between RNA methylation status and RNA half-life is observed in mouse embryo stem cells (pvalue = 2.23e-12, generalized linear model of binomial family).

Dys-regulation of RNA methylome after Simian Retrovirus Infection

Simian retrovirus (SRV) infection of Jurkat T lymphocytes (Jurkat cells) was confirmed by syncytia formation, of which the membrane of the neighboring cells fused to one another. At 10 days postinfection, the formation of syncytium was observed among the Jurkat cells incubated with SRV (**Figure 14a**). The syncytium of Jurkat cells contains multiple nuclei and its size is dramatically larger than a single cell. SRV long terminal repeats (LTRs), which are reverse-transcribed from the RNA genome during the infection, contain critical sequences necessary for the integration, synthesis, and expression of

viral DNA [1]. Therefore, the extent of SRV infection was assayed by monitoring SRV-LTR expression in Jurkat cells through quantitative realtime PCR. As shown in **Figure 14b**, the copy number of SRV-LTR was gradually increased from 2 days to 10 days postinfection and then tended to be stable afterwards. Taken together, these results indicated that SRV was able to infect Jurkat cells and the infection reached maximum level after 10 days postinfection.

In order to investigate whether SRV could replicate in Jurkat cells, the SRV virions released in the culture medium were determined by measuring viral genome copy number through quantitative realtime PCR. As shown in **Figure 14c**, the copy number of SRV genome was gradually increased during 2 days to 14 days postinfection, suggesting that SRV was able to replicate in Jurkat cells.

We then measured the RNA methylome with bisulfite sequencing. A total of 2475 m5C sites located on 517 genes are reported as differentially methylated 10 days post infection of SRV with QNB p value < 0.05. Among them, 389 sites located on 158 genes are hypo-methylated; while 2086 sites from 382 genes are hyper-methylated. A gene ontology analysis using the DAVID website suggests that the differentially methylated genes are related to virus infection, specifically, hyper-methylated genes are enriched with DNA replication (pvalue = 6.07E-5), mitotic nuclear division (pvalue = 4.37E-4), DNA replication initiation (pvalue = 3.48E-3), autophagosome assembly (pvalue = 8.54E-3), strand displacement (pvalue = 1.42E-2), double-strand break repair via homologous recombination (pvalue = 3.42E-3), etc.; while hypo-methylated genes are enriched with the following biological processes, including, negative regulation of epidermal growth factor receptor signaling pathway (pvalue = 3.24E-3), DNA damage checkpoint (pvalue = 2.54E-2), cell migration (pvalue = 1.42E-2), etc (see **Figure 14**). Similar to before, a positive correlation (0.07) is observed between RNA methylation level and expression level; however, as there are 23 genes that carry hyper and hypo-methylated sites simultaneously, it is expected that RNA m^5C carries more complicated biomolecular functions.

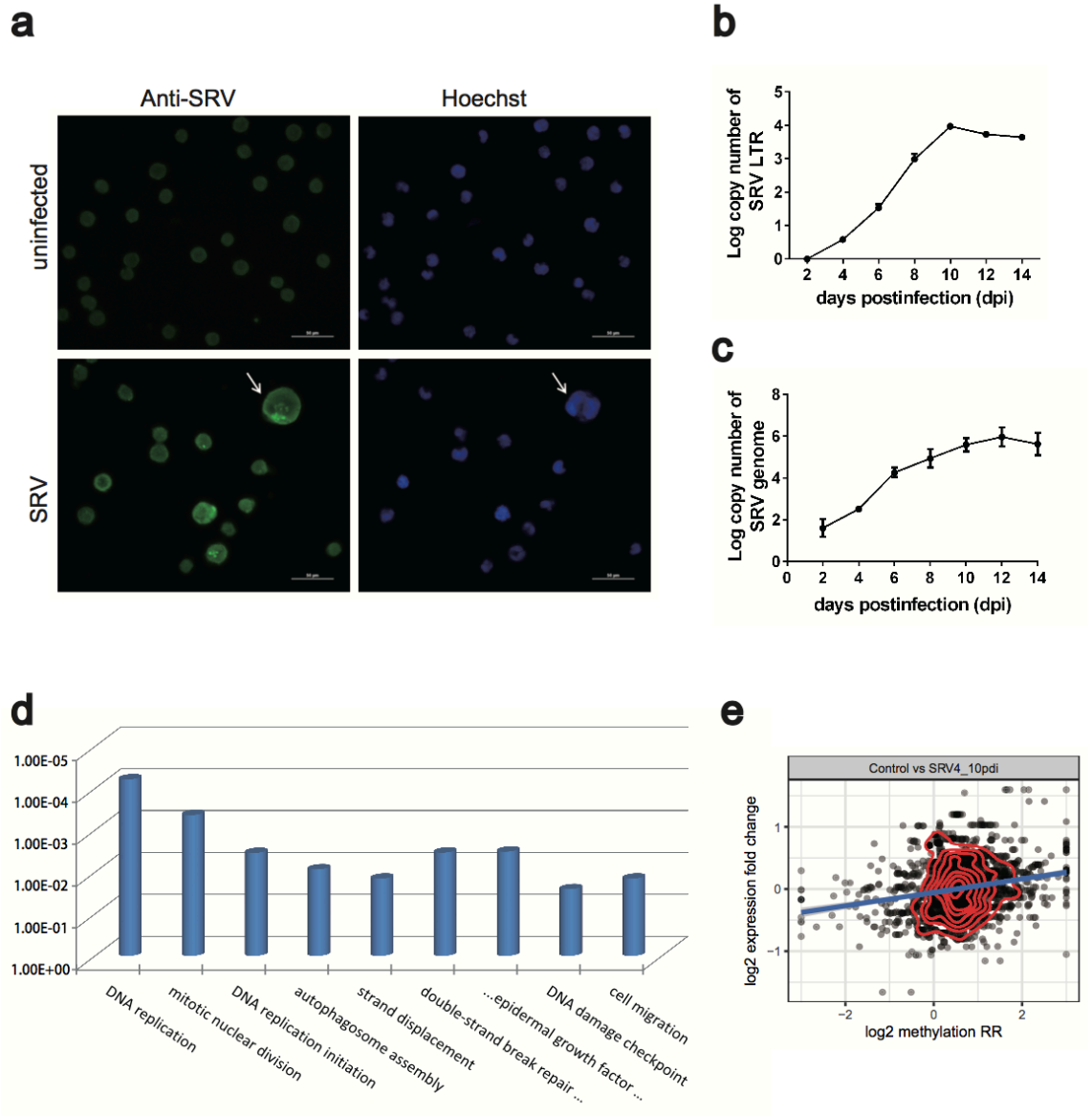


Figure 14. Dysregulation of RNA methylome after SRV infection of Jurkat Cell. (a) At 10 days postinfection, uninfected or SRV-infected Jurkat cells were stained with SRV antibodies (green). Nuclei were visualized by Hoechst staining (blue). Arrows indicate the syncytium of infected cells. Scale bar: 50μM. (b) The relative level of SRV-LTR in infected Jurkat cells was measured every two days by realtime PCR. GAPDH was used

as the internal control. The relative level of SRV-LTR at each time point was normalized to the data at 2 dpi; mean ± SD, n=3. (c) The absolute copy number of SRV genome in culture medium was measured every two days by realtime PCR. SRV-LTR and SRV genome were not detected in all uninfected cells and culture medium respectively; mean ± SD, n=3. **(d)** The differentially methylated genes are enriched with the following functions, including DNA replication (pvalue = 6.07E-5), mitotic nuclear division (pvalue = 4.37E-4), DNA replication initiation (pvalue = 3.48E-3), autophagosome assembly (pvalue = 8.54E-3), strand displacement (pvalue = 1.42E-2), double-strand break repair via homologous recombination (pvalue = 3.42E-3), etc.; while hypo-methylated genes are enriched with the following biological processes, including, negative regulation of epidermal growth factor receptor signaling pathway (pvalue = 3.24E-3), DNA damage checkpoint (pvalue = 2.54E-2), cell migration (pvalue = 1.42E-2), etc. (see **Figure 14**). **(e)** A weak but positive correlation (Pearson correlation = 0.07) is observed between RNA methylation level and expression level, which is consistent our previous result; however, as there are 23 genes that carry hyper and hypo-methylated sites simultaneously, which implies that RNA m^5C carries more complicated biomolecular functions.

1.5 Discussion and Conclusion

The distribution of m^5C methylation in mRNA has been mysterious with inconsistent evidence reported from previous studies [35, 37]. Here, we profiled the human and mouse m^5C epitranscriptome using RNA BS-Seq data in human MCF10A, human MDA468, mouse ESC, and mouse whole brain cells. To eliminate the data sample bias, we employed a rigorous quality control procedure by filtering false positive m^5C sites due to secondary structure and performed a comprehensive comparative analysis on cross-species conserved locus, cross-sample comparison of topological transcriptome distributions of m^5C, and differential m^5C analysis. Our analysis clearly shows that m^5C is enriched at 5'UTR in human and mouse cells, confirming the discovery of a few independent studies [35, 36, 47]. Additionally, an unambiguous correlated methylation pattern is observed on 5'UTRs, but not on CDS and 3'UTR, in different mouse and human

cell lines/tissues, suggesting a more complex aggregation pattern of m^5C that may be further characterized. Together, these observations strongly imply the functional relevance of m^5C RNA methylation and 5'UTR of mRNA. It is important to note that, although we failed to observe a correlated m^5C methylation pattern on CDS and 3'UTR regions of mRNA, it is still possible that such pattern may emerge on strictly matched cell lines/tissues.

When comparing the DNA and RNA methylome in matched cell lines in human and mouse, a negative correlation in the methylation level is observed on matched locus on DNA and RNA, which is quite surprising given that the methyltransferase of DNA and RNA may share strong sequence homology [53]. And this anti-correlation pattern is consistent at all four CG containing trinucleotides contexts, and ruled out the possibility of sample contamination or off-target effect, which should both lead to false positive correlation in data. It is possible that there exist a underlying biomolecular mechanism that function on matched locus of DNA and RNA in parallel to ensure their orchestrated methylation status.

Similar to DNA methylation, a clustering effect of m^5C on mRNA is also observed in both human and mouse. The local dependency, i.e., adjacent cytosine locus often exhibit similar methylation status, has been widely used in DNA methylation data analysis for more robust and accurate quantification of epigenetics status [58-60]. It is reasonable to expect that similar statistical approaches may be carried over into the field of single-base resolution RNA methylation data to enhance the analysis of bisulfite RNA methylation sequencing data. It is worth to mention that, around 30%-43% of m^5C residuals exist in pair in our results after filtering potential secondary structures that may lead to incomplete conversion and false positive m^5C sites. The number may be over- or under-estimated because of the unfiltered secondary structure, which leads to an over-estimation of the clustering effect, and structured regions excluded from the analysis, which may affect the estimation in both directions. It is necessary to develop

more sensitive unbiased approach that can eliminate the impact of RNA structure to more accurately assess the distribution of transcriptome m^5C modification.

Intriguingly, we observed a strong enrichment of m^5C methylation on mitochondrial transcripts with more than 50 folds of enrichment. Previously, it was reported that methyltransferase NSUN5 can regulate mitochondrial gene expression [55], and we speculate that RNA m^5C methylation may play a more vital regulatory role in mitochondria related biological processes.

Additionally, in order to have a glimpse of the dynamics of m^5C on mRNA, differential RNA methylation analysis was performed between breast cancer cell line MDA-MB-468 and the control cell line MCF10A, a total of 47genes are reported to be differentially methylated, including RTN4, NME2, CASP14, HSPB1, RPL11 and RPS3, which are related to apoptosis and programmed cell death. Although we showed previously that m^5C on mRNA are more likely to be linked to 5'UTR function, it is observed that the differential methylation sites between breast cancer cell line and normal control cell lines are mostly located on CDS and 3'UTR. These observations together implied a profound role of m^5C methylation on different regions of mRNA and in cancer pathology.

Interestingly, an overall positive correlation between RNA m^5C methylation and RNA expression level is observed in our mouse and human datasets, which added to the growing importance of mRNA m^5C methylation in regulating gene expression. Although the specific molecular mechanism is not yet clear, the observed positive correlation between RNA m^5C and RNA expression echoes our previous observed anti-correlation between DNA and RNA m^5C methylation from a different perspective, because it has been well established that DNA methylation is anti-correlated with RNA expression. However, as is known that the most abundant RNA modification m^6A methylation may enhance or reduce the stability of RNA molecule through interaction with different m^6A readers [14, 61] or regulates RN-protein interaction [62], it is reasonable to assume that

RNA m^5C may have versatile functionalities, and may get dominated by distinct mechanism under a specific condition.

In summary, our study presented an in-depth topological characterization of the m^5C RNA methylome in human and mouse. There are interesting patterns depicted and quantified, which call for further studies and novel biomolecular mechanisms for explanation.

1.6 Availability of data and materials

The data analysed in this study is obtained from public resources, including DNA BS-Seq data from MCF10A (GEO GSM659628), transcriptome m^5C methylation data from MCF10A, MDA468 (GSE84230), mouse embryo stem cell (ESC), and mouse whole brain (GEO GSE83432), and mouse ESC DNA methylation data (GSM1873374) [7, 36, 42].

The rBS2ndStructure R package for filtering artifacts in RNA bisulfite sequencing data is publically available at Github (https://github.com/ZW-xjtlu/rBS2ndStructure) with precomputed RNA secondary structures of mouse and human genome mm10 and hg19 for convenient processing of RNA BS-Seq data from popular genome assembly.

Chapter 2

TREW: a database for the epitranscriptome targets of RNA modification readers, writers and erasers in human, mouse and fly

2.1 Outline

Motivation: Epitranscriptome has emerged to be an important layer for gene expression regulation. With the advancements in high-throughput sequencing technology, the transcriptome-wide distribution of various RNA modifications such as m^6A, m^5C, m^1A, and ψ (pseudo-uridination) become increasingly available; however, till this day, the gene regulatory circuit at the epitranscriptome layer is still not comprehensively mapped or effectively integrated.

Result: Post-transcriptional RNA modifications are directly recognized, deposited or removed by cognate factors termed readers, writers and erasers. By integrating the information from 36 independent high-throughput sequencing experiments, an online database TREW was built to host 171,551 epitranscriptome targets of the RNA modification readers, writers and erasers predicted un- der different biological contexts in human, mouse and fly. To facilitate the query, search and browse of this database, the target information has been integrated into a genome browser and annotated by their association to genes, consensus motifs and microRNA binding sites, etc.

Availability: TREW database is available at: www.xjtlu.edu.cn/trew

2.2. Introduction

Our knowledge of the epitranscriptome has been greatly expanded with the advancements in high-throughput sequencing techniques such as m^6A-seq, miCLIP, Bisulfite-seq, Aza-IP, Ψ-seq and m^1A-seq [63, 64]. Epitranscriptome has emerged as an important layer for gene expression regulation, where a number of important biological functions are regulated through reversible RNA modifications [65]. With the efforts of bioinformatics community, the transcriptome-wide distribution of various RNA modifications is publically available from RMBase and MeTDB [66, 67]; however, till this day, the gene regulatory network at the epitranscriptome layer is still not comprehensively mapped or effectively integrated.

Post-transcriptional RNA modifications are directly regulated by protein factors including the readers, writers and erasers [68]. For RNA N6-methyl-adenosine (m^6A) modification, METTL3, WTAP, METTL14 and KIAA1429 have been identified as the components of RNA m^6A methyltransferase complex (writer) [69, 70], while FTO and ALKBH5 are discovered as m^6A demethylase (eraser) [71, 72].

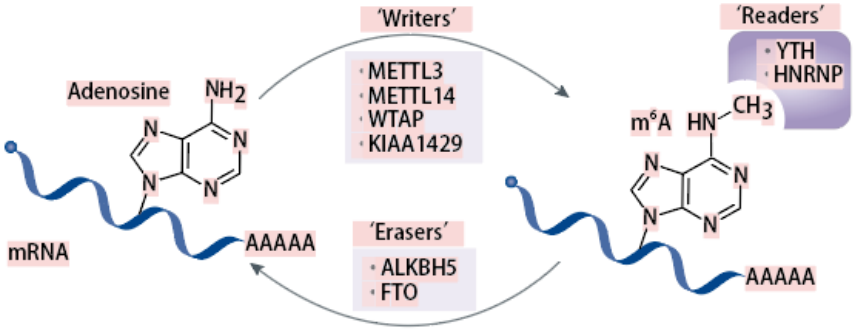


Figure. 1. Dynamic regulation of m^6A methylation. M^6A modifications on messenger RNAs are dynamically regulated by 3 groups of proteins: Writers, Erasers, and Readers. Writers are methyltransferase that could convert adenosine into m^6A. Erasers are demethylase that could convert m^6A back into adenosine. Readers could bind to m^6A

modified mRNAs with m^6A binding domains, and they will regulate the m^6A modified mRNAs by recruiting downstream effectors.

YTH family proteins YTHDF1, YTHDF2, and YTHDF3 can bind with the m6A modified nucleotides (reader) and regulate translation [73]. Besides, such protein regulators are also identified for other types of RNA modifications, e.g., m^1A can be erased by ALKBH3 [64], while Ψ is deposited by PSUS1.

It is important to note that, the direct regulators of the epitranscriptome do not target aimlessly but exhibit substrate specificity. E.g., the m^6A demethylase FTO and ALKBH5 are able to discriminate substrates with very similar nucleotide sequences [74]; while after the perturbation of WTAP, METTL14 and KIAA1429, the m^6A methyl-transferase complex exhibited altered substrate preference [70]. However, despite there exists a number of published experiments studying the context-specific regulatory effects of RNA modification readers, erasers and writers, such information has not been integrated into existing databases of RNA modifications such as RMBase [67], Met-DB [66] and MODOMICS (Machnicka, et al., 2012) [75] [75]. Hence, we developed TREW, which denotes the database for the epitranscriptome targets of RNA modification readers, erasers and writers. TREW includes the targets information of RNA modification readers, erasers and writers for multiple modification types in human, mouse and fly collected from different biological contexts, presenting an important collection of the regulatory circuit at the epitranscriptome layer for current researchers in RNA epigenetics.

2.3. TREW database

A total of 171,551 RNA modifications sites that preferentially targeted by a specific protein regulator (RNA modification reader, writer or eraser) in human, mouse and fly

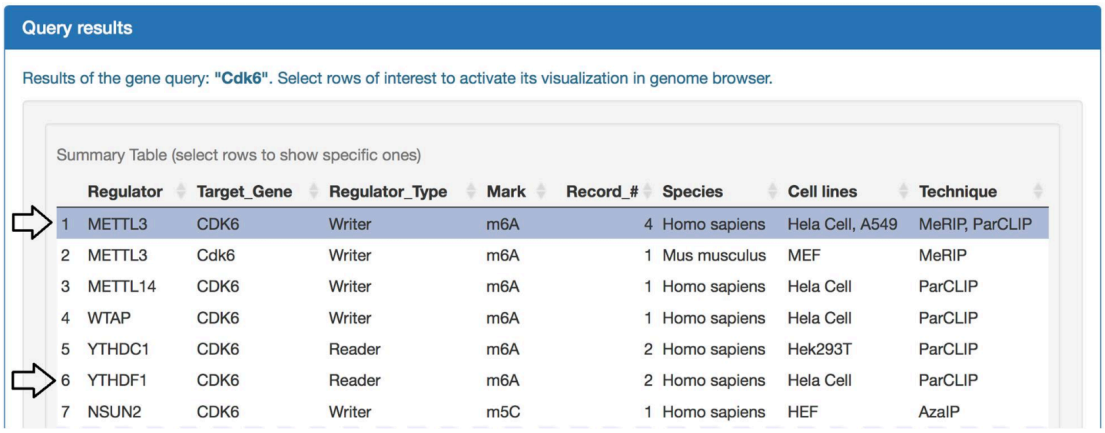
are collected from 36 sets of high- throughput sequencing experiments in 16 studies. The records are predicted from different techniques (ParCLIP, Aza-IP, m^6A-seq, Bisulfite-seq, Ψ-seq and m^1A-seq) with corresponding data processing pipelines and cover 5 major RNA modification types (m^6A, m^5C, m^1A, hmrC and ψ). The collected target sites are further annotated with transcript regions (5'UTR, CDS, 3'UTR, stop codon, transcription starting sites, miRNA target site) and RNA types (tRNA, mRNA, lncRNA, sncRNA). The annotated data are then imported into an SQL-lite database (see supplementary materials for more information), based on which the TREW web server is created under Shiny web framework and embedded with JBrowse genome browser [76].

Three query modes are supported in TREW database for different purposes, including:

- **General Query**: This mode allows the query based on target gene or the regulator. A summary and the detailed records from each species are returned (see Figure. 2), and they are linked to a genome browser for visualization purpose.

- **Orthologous Query**: The query under this mode is similar to species-specific query, except that the records from different species are merged together, which inexplicitly assumes that the regulator of RNA modification is also conserved across different species, just like the modification site itself [77, 78]. This query mode may be particularly useful for integrated analysis of closely related species, such as human and mouse.

- **Browsers**: this mode allows the users to directly browse and search the records of interests within a genome browser interface.



	Regulator	Target_Gene	Regulator_Type	Mark	Record_#	Species	Cell lines	Technique
1	METTL3	CDK6	Writer	m6A	4	Homo sapiens	Hela Cell, A549	MeRIP, ParCLIP
2	METTL3	Cdk6	Writer	m6A	1	Mus musculus	MEF	MeRIP
3	METTL14	CDK6	Writer	m6A	1	Homo sapiens	Hela Cell	ParCLIP
4	WTAP	CDK6	Writer	m6A	1	Homo sapiens	Hela Cell	ParCLIP
5	YTHDC1	CDK6	Reader	m6A	2	Homo sapiens	Hek293T	ParCLIP
6	YTHDF1	CDK6	Reader	m6A	2	Homo sapiens	Hela Cell	ParCLIP
7	NSUN2	CDK6	Writer	m5C	1	Homo sapiens	HEF	AzaIP

Figure. 2. Query result of CDK6. A summary is returned with records from different species. CDK6 is predicted from MeRIP-seq and ParCLIP to be preferentially methylated by METTL3 in 2 different human cell lines. It is also targeted by m^6A reader YTHDF1 in 2 ParCLIP samples performed in human Hela cell line. Detailed information will be provided when the relevant summary record is selected.

2.4 Raw Data Collection

The transcriptome wide m^6A methylation is mostly detected by technique of m^6A-Seq[79]. In m^6A-Seq, purified RNAs are fragmented into ~100 nt fragments. Then, the fragments undergo Immunoprecipitation (IP) with anti-m6A antibodies. After RNA-seq sequencing, the reads are aligned to the genome; The IP group and input control group (the fragments without IP) are compared to infer the genomic locations containing m6A methylation.

TREW database collects the binding sites of m^6A methyltransferases (METTL3, WTAP, METTL14 and KIAA1429), demethylases (FTO and ALKBH5) and readers (YTH family proteins). To determine the target sites, ParCLIP-seq data were retrieved directly from original publications, where the raw data were processed with Trim Galore and FASTX-Toolkit (v0.0.13) for quality control, and then aligned to human hg19 or mouse mm10

reference genome respectively with Tophat2[80]. Also, differential m^6A analysis was performed with exomePeak and QNB packages[50] under the default setting on MeRIP-seq data of m^6A methylase or demethylase perturbation. The significant differential m^6A peaks after perturbation were determined to the target peaks. Please refer to the Supplementary File 1 for the detailed documentation of the data sources, data processing pipelines, and app interface instructions for TREW database.

2.5 Analysis of Data Consistency

17 m^6A-Seq datasets from TREW were selected for the evaluation of target consistency, all of them have experimental design to study the m^6A changes after the knock down or over expression of m^6A regulators. The raw reads are trimmed and aligned with trim_galoure, fastx_trimmer, tophat2 [80], and hisat2[81]. The aligned reads are counted by 101 nt bins centered by m^6A annotations of single based resolution.

The annotations used for m^6A-Seq quantifications consist of m^6A sites from RMBase2 [67], TSS Adenine of Ref-Seq, and 6 miCLIP datasets (A technique that can detect m^6A in single based resolution). The count results on mouse (mm10) are lifted over to human (hg19). Roughly 30% of the annotations are conserved between human and mouse after Liftover. The following analysis is performed on 146310 conserved methylation sites between human and mouse.

The methylation and differential methylation (DM) analysis is conducted in DeSeq2 [51]. For samples with biological replicates, the cut-off was set at adjusted p values < 0.05, and for samples without biological replicates, the cut-off was set at p values < 0.05. If the reported methylation sites have number < 10000 for a given dataset, the extra sites are reported by order of p values until the number reached 10000. For differential methylation, the minimum number reported was set at 5000. For purpose of control,

the reversed methylated sites and reversed differentially methylated sites are reported using the same criteria.

Results:

Quality control

The quality of the data is examined by 2 methods: 1. the distribution of exon length of the methylated sites, and it is expected that the highly methylated sites are enriched on long exons; 2. the distribution of the methylated sites on transcript coordinate, and it is expected that the m6A sites are enriched around stop codon. For all the datasets, the significant methylated sites are enriched in long exons and stop codons compared with the reversed methylated sites (See Supplementary File 2).

The same metrics are applied to examine the quality of the differential methylation. All the writers show selectivity toward long exons and stop codons. This confirms with the previous study [69]. However, the erasers show different patterns of genomic features compared with writers. Most of the eraser targeted sites cannot demonstrate clear enrichment on long exons and stop codons compared with the reversed DM sites. Although the annotation sites used for quantification are high confidence m6A sites, the enrichment of erasers targets on long exons and stop codons are generally weak. Hence, the selectivity of erasers may not be governed by the genomic features that are important to predict high methylation levels.

- Visualizing Similarities

		Data Set 1	
		Expected DM	Reversed DM
Data Set 2	Expected DM	a	b
	Reversed DM	c	d

$$OR = \frac{ad}{bc} \qquad dist = \sqrt{\frac{c+d}{a+b+c}}$$

The similarities between the datasets are evaluated by a 2 by 2 contingency table. Odds ratio (OR) is used to measure the similarities on the contingency table, and the square root of unsymmetrical distance is used to measure the dissimilarities.

The log2 OR is visualized by heat map (Figure 3) with maximum values truncated at 5. the hierarchical clustering based on the entries of the matrix shows that the writers and erasers can form 2 clusters. We further verify the clustering results with MDS (multiple dimensional scaling) (Figure 4); the distances between the dots on the MDS plot are approximately equal to the unsymmetrical distances. However, we cannot observe the clustering effect within the same writer regulator. The results may suggest that the writer proteins are part of a big methyl-transferase protein complex, and they might function as a single unit. The eraser proteins also form a single cluster, and it indicates that Fto and alkbh5 show similar target specificities on m6A.

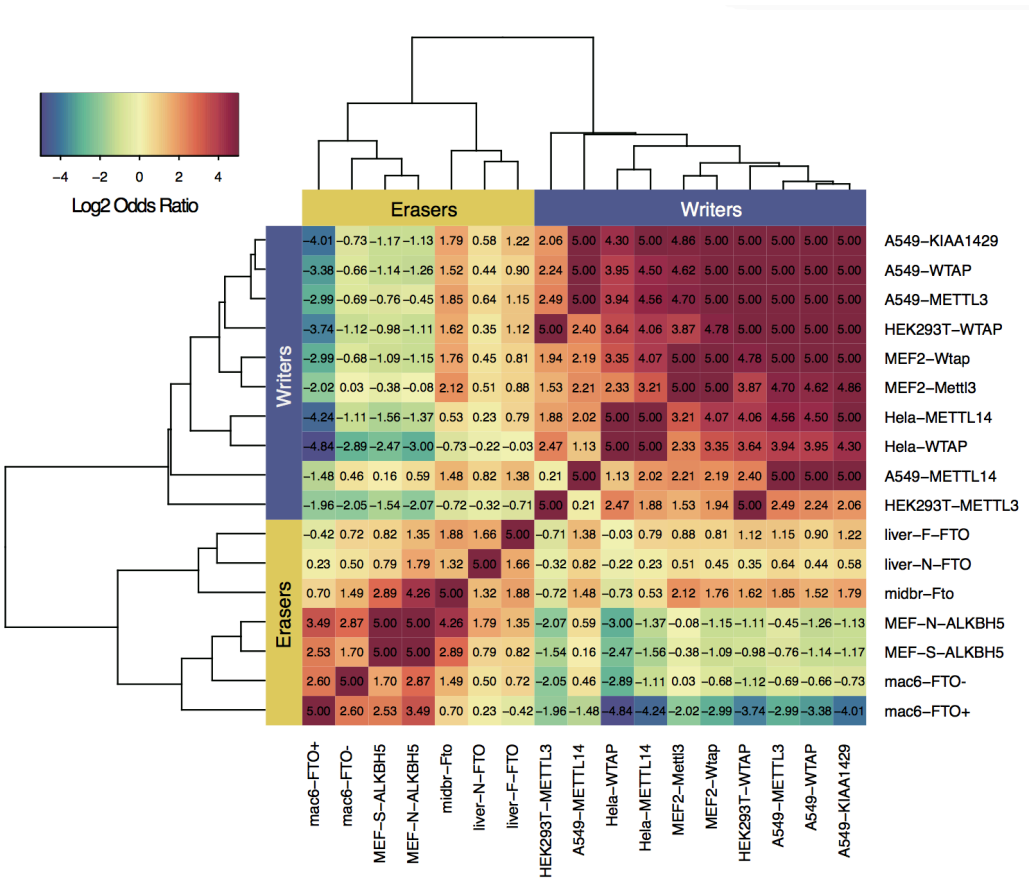


Figure 3. Heat Map of Odds Ratio Between Samples. The entries in the matrix are Log2 Odds Ratio associations of the significantly expected DM sites and the significantly inverse DM sites between different perturbed MeRIP-Seq data sets. The association values are truncated at 5. The hierarchical clustering results based on the OR matrix are labelled at the margin. The dendrogram indicates that the data sets could be divided into 2 clusters, where the clustering partition can be clearly explained by the distinction of Writers and Eraser regulators. In addition, the writer and eraser proteins show higher within group association values than between group association.

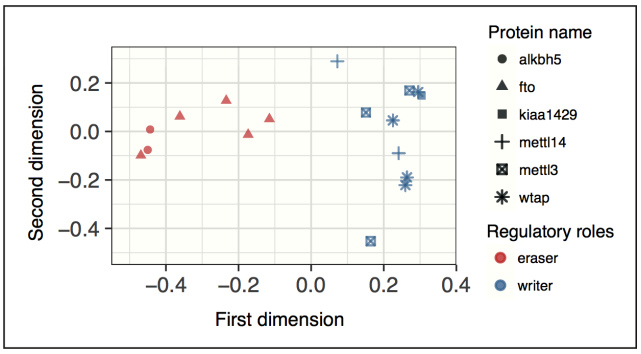


Figure 4. MDS Plot of Asymmetrical Distances Between Samples. MDS is conducted on the square root of asymmetrical distances calculated from the same contingency table of the OR association values. The plot based on the 2 scaling coordinates indicate that the expected differential methylation profiles are indeed demonstrated consistency within the writer and eraser groups.

To further evaluate the statistical significance of the clustering, we fit a linear regression model on log2 OR between different samples. The model selection is conducted with the R package BAS. Based on the Bonferroni adjusted p values of Wald test. We identified 3 significant predictors with adjusted $p < 0.05$ to explain the OR variabilities; the significant predictors are writers (padj = 2.44e-23), erasers (padj = 5.06e-05), and cell lines (padj = 0.0001). The complete linear model results and model diagnosis are included in Figure 5.

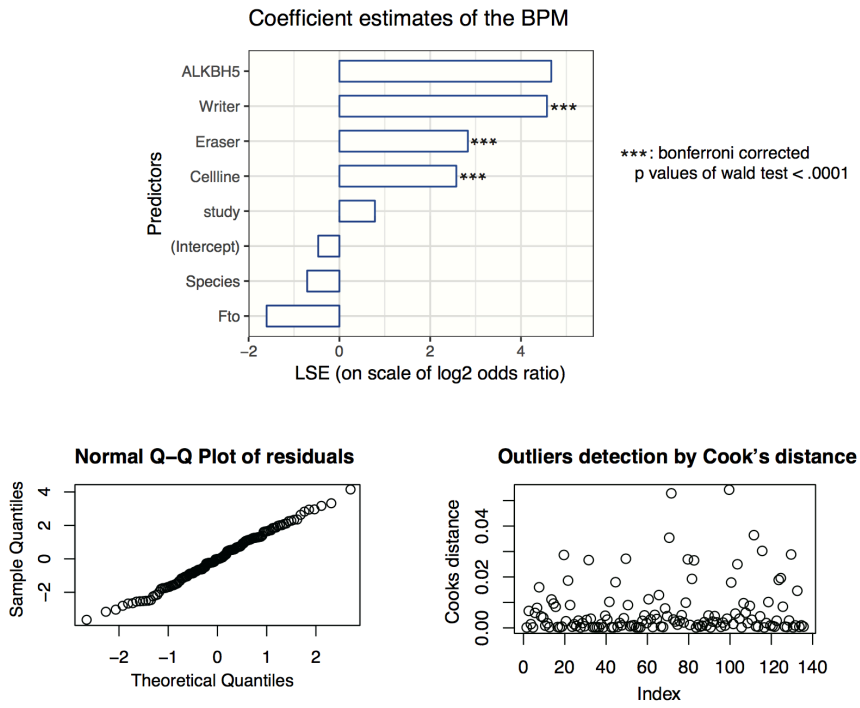


Figure 5. Linear Regression Model on log2 OR values. A linear regression model is fitted on log2 OR matrix using predictors of the 7 dummy variables. The variables are indicator of weather the 2 data sets belong to the same regulator group, cell lines, study, or species. The coefficient estimates are significantly positive for variables of Writer, Eraser, and Cell lines, which indicates the statistical significant contribution to the target consistency. Also, the diagnosis plot indicates the assumption of linear regression model is satisfied.

To explain the details of the specificities, we construct a logistic regression model with 25 predictors of transcriptomic features (Table 1). The features are used to classify sites into consistently targeted sites and consistently not targeted sites for either writers or erasers. The consistently targeted sites are defined by the sites that are targeted by at least 2 datasets; the consistently un-targeted sites are defined by the sites that are reversely targeted in at least 2 datasets. The contradictory sites that mapped to both groups are removed from our analysis.

	Predictor names	Definition
1	**Protein coding**	Sites mapped to exons of coding genes.
2	**CDS/5'UTR/3'UTR**	Sites mapped to coding sequence/5'UTR/3'UTR.
3	**Long exon**	Sites mapped to exons longer than 400bp.
4	**Last exon**	Sites mapped to the last exon in a transcript.
5	**Introns**	Sites mapped to introns.
6	**Relative positioning**	The relative position of sites on spliced transcript.
7	**TSS**	100bp downstream of the TSS
8	**Stop codon**	400bp centered by stop codon.
9	**Start codon**	200bp centered by start codon.
10	**DRACH/BCA**	Sites on DRACH/BCA consensus motif.
11	**Struc Sterm/loop**	Loop/sterm of thermal stable RNA structures
12	**PC score1/200**	PhasCons Score in bins of 1nt and 200 nt.
13	**HK genes**	Sites on housekeeping genes.
14	**Isoform numbers**	Number of transcript iso-forms in the gene.
15	**TREW overlapping**	overlapping with PARCLIP datasets in TREW of HNRNPC, YTHDF2, YTHDF1, and YTHDC1

Table 1. The Feature Names and Their Definitions. Genomic related features are constructed to predict the regulator specificities of consistent writer and eraser targets. The genomic features are created based on the m^6A topological significance demonstrated by the previous studies[69, 78, 82].

The final model is selected by BAS. A robust coefficient prior and a beta binomial (1, 1) model prior are selected. After 20000 iterations of MCMC sampling, the final model is determined by the best predictive model (BPC) method.

The bar plot in Figure 6 indicates the logit (log odds) estimates and the z statistics of wald test in the BPMs. The more positive or negative of the logit, the more enlargement or shrinkage it can made toward the probabilities of the sites being specifically targeted. The higher the z statistics in absolute value, the more statistically significant the predictor is.

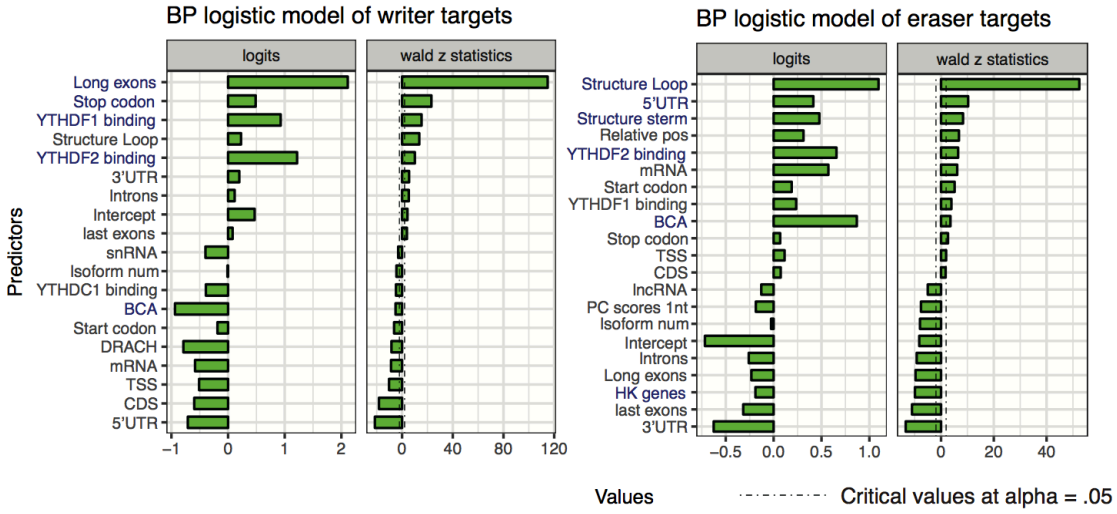


Figure 6. Logistic Regression Models using Genomic Features. The plots above demonstrate the coefficients or logits estimates of the logistic regression models using genomic features. The response variables used are consistent targets of writers and erasers. Both of the model prdictors are selected using Bayesian model selection methods under the best predictive model criterion. The resulting coefficient estimates are all statistical significant under Wald tests after Bonferroni correction. The targets of writers and erasers show substantial differences on their profile of associations with genomic features.

From the logistic model, we can infer that the selectivities of the writers and erasers are explained by different transcript features. Specifically, writer prefer long exons and stop codon, which help us to explain most of the deviances in the data. Erasers prefer looped RNA 2ndary structure, 5'UTR, and BCA motif. Additionally, both of the writer and eraser are highly associated with YTHDF2 binding site. This result suggests that both of the target selectivities are enriched on true m6A modification sites rather than potential false-positive sites.

We confirmed that the writers and erasers have different specificities, and the specificities are evolutionary conserved. We also observed that the eraser specificities are highly structural driven, and the writer specificities are exon length and stop codon driven. The structural dependency of erasers was also suggested by a wet experiment study [74], and our results confirmed the picture of the regulator specificities from a computational approach. We present an alternative but more comprehensive approach to infer the specific structural, topological, and functional factors that drive the specificities of eraser and writer proteins.

2.6 High accurate m^{6}A site prediction with hand crafted genomic features.

Most existing RNA modification site prediction algorithms use exclusively sequence-based features; however, the sequence-derived features alone may not fully capture the attributes of RNA modification topology. Hence, we generated 45 additional genomic features that may contribute to the prediction. Genomic Features 1-13 are dummy variable features indicating whether the adenosine sites fall within the transcript regions that satisfy certain topological properties. All of the features in this category are generated by GenomicFeatures R/Bioconductor package using the transcript annotations hg19 TxDb package. To remove the ambiguity caused by transcript isoforms, only the primary (longest) transcripts of each gene were kept for the extraction of the transcript sub-regions. Genomic Features 14-16 are real valued features defining the relative

position of the transcript regions. Specifically, it is defined by distance from the adenine to the 5' end divided by the total width of the region; for the adenosine sites not belong to the region, the position features are set 0. Genomic features 17 to 19 represent the length of the transcript region containing the modification site. The values are also set to 0 for sites that not belong to the region. Features 20-22 captures the nucleotide distance of adenine sites toward the 5'end/3'end of the splicing junctions. Additionally, the distance to the nearest neighboring m^6A sites in the training data is generated to measure the clustering effect of the m^6A RNA modification. Features 23-26 represent the evolutionary conservation score of the adenosine sites and its flanking regions; 2 metrics of nucleotide conservation, Phast-Cons score [83] and the fitness consequence scores [84], are used to measure the conservation level of the underlying nucleotide sequence. Features 27 and 28 represent the RNA secondary structures around the adenine site, the RNA secondary structures are predicted on the exon regions using RNAfold in Vienna RNA package [85]. At last, features 29-35 are the properties of the genes or transcripts containing the m^6A sites, such as the miRNA target genes and the housekeeping genes. The annotation of microRNA target sites are from miRanda [86] and TargetScan [87].

Details about the genomic features used in the prediction model.

Together, 69 features were crafted to predict the m6A sites on both the full transcript regions and exons, and 35 of them were selected in the final prediction model using the method of recursive feature elimination. The 35 features used in the final model are:

Table 7. Genomic Features Selected in the Whistle classifier

ID	Name	Description	Note
1	UTR5	5' UTR	Dummy variables indicating whether the site is overlapped to the topological region on the major RNA transcript.
2	UTR3	3' UTR	
3	Stop_codons	stop codons flanked by 100bp	
4	Start_codons	start codons flanked by 100bp	
5	TSS	downstream 100bp of TSS	
6	TSS_A	downstream 100bp of TSS on A	
7	exon_stop	exons containing stop codons	
8	alternative_exon	alternative exons	
9	constitutive_exon	constitutive exons	
10	internal_exon	Internal exons	

11	long_exon	long exons (exon length >= 400bp)	
12	last_exon_400bp	5’ 400bp of the last exons [88]	
13	last_exon_sc400	5’ 400bp of the last exons containing stop codons [88]	
14	pos_UTR5	relative position on 5'UTR	Relative position on the region
15	pos_UTR3	relative position on 3'UTR	
16	pos_exons	relative position on exon	
17	length_UTR5	5'UTR length	The region length in bp.
18	length_UTR3	3'UTR length	
19	length_gene_ex	mature transcript length	
20	dist_sj_5_p2000	distance to the 5' splicing junction	Nucleotide distances toward the splicing junctions or the nearest neighboring sites.
21	dist_sj_3_p2000	distance to the 3' splicing junction	
22	dist_nearest_p200	distance to the closest neighbor truncated at 2000bp	
23	PC_1bp	phastCons scores of the nucleotide [83]	Scores related to evolutionary conservation
24	PC_101bp	average phastCons scores within the flanking 50bp region [83]	
25	FC_1bp	fitCons scores of the nucleotide [84]	
26	FC_101bp	average fitCons scores within the flanking 50bp region [84]	
27	struc_hybridize	Predicted RNA hybridized region [89]	RNA secondary structures
28	struc_loop	Predicted RNA loop region [89]	
29	sncRNA	sncRNA	Attributes of the genes or transcripts
30	lncRNA	lncRNA	
31	HK_genes	housekeeping genes [90]	
32	miR_targeted_genes	miRNA targeted genes [91]	
33	HNRNPC_eCLIP	eCLIP data of HNRNPC RNA binding sites [92]	RNA annotations related to m6A biology.
34	Verified_miRtargets	miRNA targeted sites verified by experiment[93]	
35	TargetScan	Predicted miRNA targeted sites by TargetScan[94]	

Machine learning approach used for m^{6}A site prediction

SVM is one of the most widely used machine learning algorithms in computational biology, which has been previously used for mammalian microRNA target prediction [95], protein kinase-specific phosphorylation sites prediction [96] and mammalian m^{6}A modification sites prediction [97, 98]. In this project, we used an R language interface of LIBSVM [99] to construct the SVM-based m^{6}A site predictors. The radial basis function was chosen as the kernel function, and other parameters were set at the default.

Performance evaluation of m^{6}A site prediction

For the SVM classifier, a 5-fold cross-validation is employed on the training datasets for model selection purpose, and the final performance of the predictor is measured on the independent testing dataset. The ROC (receiver operating characteristic) curve

(sensitivity against 1-specificity) is used to measure the prediction performance under different decision thresholds, and the area under ROC (AUC) was calculated as the main performance evaluation metrics.

When evaluating the accuracy of m^{6}A site information stored in existing epitranscriptome m^{6}A site databases MeT-DB Version 2 and RMBase Version 2, the reliability is determined by the number of experiments that support the existence of a specific m^{6}A site, based on which the AUC can be calculated. In addition, the sensitivity (Sn), specificity (Sp) and Matthews correlation coefficient (MCC) were calculated to measure the performance of predictor:

$$Sn = \frac{TP}{TP + FN} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{5}$$

where, TP , TN , FP and FN represent true positive, true negative, false positive, and false negative, respectively.

The performance of the proposed m^{6}A predictors are then evaluated using independent datasets and compared with competing approaches. As shown in **Table 8**. By combining additional genome-derived features, the performances of our approach are substantially higher on all the tested conditions than MethyRNA and SRAMP, which relies on only information extracted from sequences. The WHISTLE achieved AUCs of 0.960 and 0.895 under the full transcript and mature mRNA modes, respectively, representing a major improvement compared with MethyRNA (0.83 and 0.772) and SRAMP (0.835 and 0.794).

Table 8. Performance of m^{6}A Site Prediction

Model	Method	Performance on Independent Dataset (AUC)						Average AUC
		A549	CD8T	Hela	HEK293 (sysy)	HEK293 (abacm)	MOLM13	
Full Transcript	WHISTLE	0.973	0.947	0.962	0.954	0.976	0.947	0.960
	MethyRNA*	0.844	0.837	0.781	0.884	0.824	0.807	0.830
	SRAMP	0.871	0.859	0.791	0.900*	0.861	0.794	0.835
Mature mRNA	WHISTLE	0.923	0.922	0.911	0.904	0.857	0.856	0.895
	MethyRNA*	0.796	0.786	0.724	0.825	0.759	0.742	0.772
	SRAMP	0.833	0.820	0.738	0.884*	0.824	0.754	0.794

We thus performed a whole transcriptome prediction of m^{6}A RNA methylation sites in human to generate a map of human m^{6}A epitranscriptome using our proposed WHISTLE approach. Our predicted map is of substantial higher accuracy (average AUC of 0.960 and 0.895) compared with existing epitranscriptome databases MeT-DB (average AUC of 0.833 and 0.782) and RMBase (average AUC of 0.822 and 0.775) when evaluated on independent base-resolution datasets under full transcript and mature mRNA mode, respectively.

Chapter3.

gcepc: R package to conduct GC content bias aware exonic peak calling and quantification in meRIP-Seq

3.1 Outline

MeRIP-Seq represents the most popular type of high through-put assay which are widely used in RNA epitranscriptomic research. In MeRIP-Seq, the locations of RNA modifications under a given cellular condition are often inferred from the statistical enrichment of the reads coverage in IP (immune-precipitated) samples over the input control samples. However, similar to RNA-Seq, as a type of 2nd generation sequencing, we observed that substantial technical variation is arised from quantification in MeRIP-Seq experiments due to the GC content bias induced during the PCR amplification process of the sequenced fragments. In addition, the GC content biases are usually correlated with the laboratory conditions, leading to systematic batch related errors in peak calling and modification level quantifications. gcepc is a novel software package which employs advanced statistical models to account for the GC content bias and biological variations. Compared with exomePeak, gcepc could significantly improve the accuracy of peak calling and modification level quantification while reducing the effect of batch variation under different laboratory conditions. Moreover, gcepc supports various meaningful functionality such as differential analysis and modification quantification.

R package **gcepc** is available at: https://github.com/ZW-xjtlu/exomePeak2

3.2. Introduction

Messenger RNA modifications represents a layer of post-transcriptional regulation that plays a crucial role in deciphering the precise mechanism for mRNA turn over, translation, and subcellular localization[100]. With the rapid development of epitranscriptomics, large amounts of high throughput assays are generated by different laboratories to measure the RNA modification profiles under different cellular conditions. Among the published datasets, MeRIP-Seq [101] (Methylated RNA immunoprecipitation sequencing) is the most commonly used HTP technique to assess the location and density of RNA modifications on messenger RNA transcripts.

As a RNA-Seq based HTP experiment, MeRIP-Seq extracts the ploy-A RNA transcripts from cell. The RNA transcripts are firstly fragmented into small RNA segments of 50 – 100 bp in lengths, and the RNA segments are separated into 2 groups: one is the IP group which is immune-precipitated by modification specific antibody; the other is the input group which is the control smaple to account for the background transcript expression level. The fragmented RNAs containing the RNA modification will be enriched by the anti-modification antibodies and therefore being over represented in the IP samples. The enrichment on reads abundance of IP over input samples will reveal the location and intensities of RNA modification under a given transcript region.

The modification sites are often identified by peak calling algorithms from the MeRIP-seq data. The most popular peak calling algorithms in MeRIP-seq are Mayer's method[102], MACS[103], and exomePeak[79]. The Mayer's method employs a fisher's exact test to examine the association between IP&input labels with the particular genomic regions. The reads counts in Mayer's method are derived from the sliding windows generated on exonic regions of the RNA transcripts. MACS is a peak calling software developed for CHIP-Seq data. It applies a Poisson statistical model with a flexible background parameter estimator to identify the targetd protein binding sites on the whole genome. exomePeak is a R package which conducts MeRIP-Seq peak calling on exons, the corresponding inference method for RNA modification targeted sites is

the exact Poisson test (c-test) on the differences between 2 Poisson means of the IP and input samples.

All the previous methods mentioned above neglect the feature specific GC content biases which are prominent in RNA-Seq related assays. In addition, the classical methods are unable to incorporate the over-dispersed variation which is common in mesarsures of gene expression levels among biological replicates. To address these important problems, gcepc employs the generalized linear model (GLM) of negative binomial (NB) family with a feature specific offset which takes the GC content bias into consideration. Under the default settings, GLM of NB used in gcepc inherents the models developed by DESeq2[51] which applies a regularized estimation method on the over-dispersion parameter using counts of all the features under a given condition.

3.2. The gcepc Peak Calling Pipeline

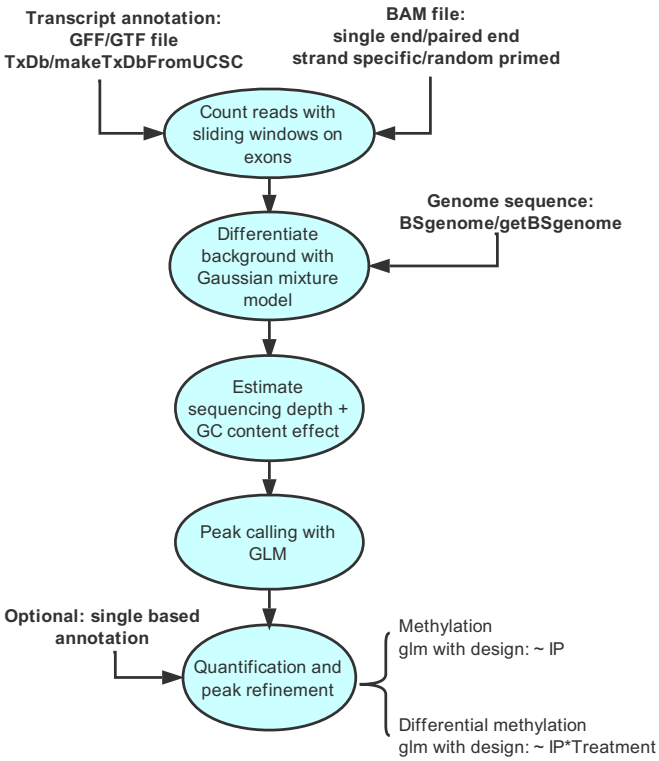


Figure1. The peak calling pipe-line of gcepc, the input for gcepc are the BAM files, genomic sequences, and transcript annotations. During Peak Calling, gcepc will first estimate the GC content offsets on user defined features, which could be the background bins that are searched by a GMM on quantities of log2 IP/input fold changes. The peak calling is then conducted on bins with GC content correction using GLM of NB. A second round of quantification will be conducted on merged bins to estimate the (differential) modification statistics by DESeq2.

As is shown in **Figure1,** gcepc takes the aligned reads of IP and input samples in BAM format as inputs. Given the transcript annotation files in either GFF or TxDb formats provided by the user, a sliding window will be generated on exon regions with the widths equalling to the size of antibody binding lengths (default set to be 25bps). The 5'POS of the reads are counted on bins flanked by the length of $(fragment\ length - \ binding\ length)$ and the step size of the sliding window is set to be equal to the binding length.

After collecting the read count statistic on the bins, the sequencing depth will be estimated on the bins using the robust sequencing depth estimator introduced in DESeq[104] followed by the estimation of the GC content linear effect which is estimated on each sample for bins with the average reads count > 50 across biological replicates. The filter for bins with low variance will be applied on IP and input separately. For each sample, the bin specific offsets of GC content bias are the centred fitted values estimated by a median regression between GC content and log reads count, this approach is firstly introduced in the CQN[105] method.

In order to effectively estimate the technical effect of GC content without the confounding between the biological association of GC and modification signal, the scope for the GC linear effect estimation can be limited to the background bins (i.e. the bins without modification signals). The background region is by default identified using a Gaussian mixture model (GMM) on the log2 IP/input enrichment ratio across bins with average reads count > 50. Other methods for the background identification are also implemented in gcepc such as using the prior information of the m^{6}A modification topology.

$$K_{i,j} \mid \rho(j) \sim NB(\mu_{i,j},\ \alpha_{i,j})$$
$$\log(\mu_{i,j}) = \beta_{0,i} + \beta_{1,i}\mathrm{I}(\rho(j) = \mathrm{IP})) + t_{i,j}$$
$$\hat{f}_{i,j} = QR(y = \log(K_{i,j}); x = X_i)$$
$$t_{i,j} \equiv \exp[\hat{f}_{i,j}(X_i) + \log(\hat{m}_j) - \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{i,j}(X_i)]$$

The Statistical Model involved in gcepc peak calling. The reads count $K_{i,j}$ in window i and sample j conditioning on the IP treatment of sample j is modelled by a negative binomial distribution with mean $\mu_{i,j}$ and over dispersion parameter $\alpha_{i,j}$. The regression equation of the conditional mean is modelled by a design with dummy variable of sample j being the input sample. An additional offset variable $t_{i,j}$is included in the regression equation. The offset parameter is defined by the last 2 lines of equations: the conditional specific GC content linear effect $f_{i,j}$ is estimated using a median regression (QR with q = 0.5) on the logarithm of reads count for bins with average count > 50. Natural cubic splines with 5 knots are applied for the feature expansion. The final offset is the sum of the logarithm of the sequencing depth

estimate and the centred GC content linear effect (for model identifiability). The peak calling is conducted using the 2-sided Wald test on the GLM coefficient $\beta_{1,i}$.

The modification status on bins are determined using a Wald test on the GLM of NB with the indicator variable of IP samples as the regression covariate. By default, the GLM used in gcepc is the GLM of NB implemented in DESeq2, which has a regularized estimation on the over-dispersion parameter. The offset of the GLM in peak calling is set to be the sum of the sequencing depth and the centered GC content linear effect estimates.

After the inference of the significantly modified bins, the bins which have p values less than 1×10^{-5} are regarded as the modification positive bins. The positive bins are merged into modification peaks, and only the peaks longer than $2 \times$binding lengths are kept in the downstream analysis. The count statistics are re-quantified on the merged modification peaks, and the same GLMs in DESeq2 are fitted on peaks to report the final log2 IP/input fold changes as well as other peak statistics. The GLM offsets used in peak quantification are re-calculated using the GC content of the underlying sequence covered by the peak.

3.3. Additional Functionalities Supported by Gcepc

- **Differential analysis**

Experimental design containing the differential analysis of MeRIP-Seq data is common among the published epitranscriptomic studies [69, 82]. Comparison between multiple conditions (sometimes has the perturbation of writers/erasers) are usually conducted in methods lacking model based rigoriousity. gcepc is the first package that supports the differential analysis of MeRIP-Seq data while adjusting the GC content bias.

$$\log(\mu_{i,j}) = \beta_{0,i} + \beta_{1,i}\mathrm{I}(\rho(j) = \mathrm{IP})) + \beta_{2,i}\mathrm{I}(\rho(j) = Treatment)) + \beta_{3,i}\mathrm{I}(\rho(j) = \mathrm{IP\&Treatment})) + t_{i,j}$$

The differential analysis is conducted after a standard peak calling pipeline while interactive GLMs using the above regression equation are fitted on the modification peaks, the peaks in differential mode are called using the pooled samples of the

treatment and control condition. The coefficient estimates of the interactive terms $\beta_{3,i}$ are the log2 Fold Change estimate of the differential modification. A 2-sided Wald test is applied on the differential log2FC to identify the significantly differentially modified RNA modification targets.

- **Quantification on Single Based RNA Modification Annotation**

gcepc supports the modification quantification and differential modification analysis on single based modification annotation. The modification sites with single based resolution can provide a more accurate mapping of modification locations compared with the peaks which report the regions of RNA modification rather than the specific sites.

Some of the datasets in epitranscriptomics have a single based resolution, such as the data generated by the m6A-CLIP-Seq or m6A-miCLIP-Seq techniques. Reads count on the single based modification sites could also provide a more accurate and consistent quantification on MeRIP-Seq experiments by eliminating of the technical variation introduced by the differences in the peak lengths.

The 5'POS of the reads are count into the single based modification sites and are flanked by the width $(fragment\ length - \ binding\ length/2)$. The methods of quantification and differential analysis used are consistent with the standard pipeline. In other word, gcepc will treat the single based modification sites as modification peaks in the downstream analysis.

3.4. The GC Conetent Biases Observed in MeRIP-Seq Experiments

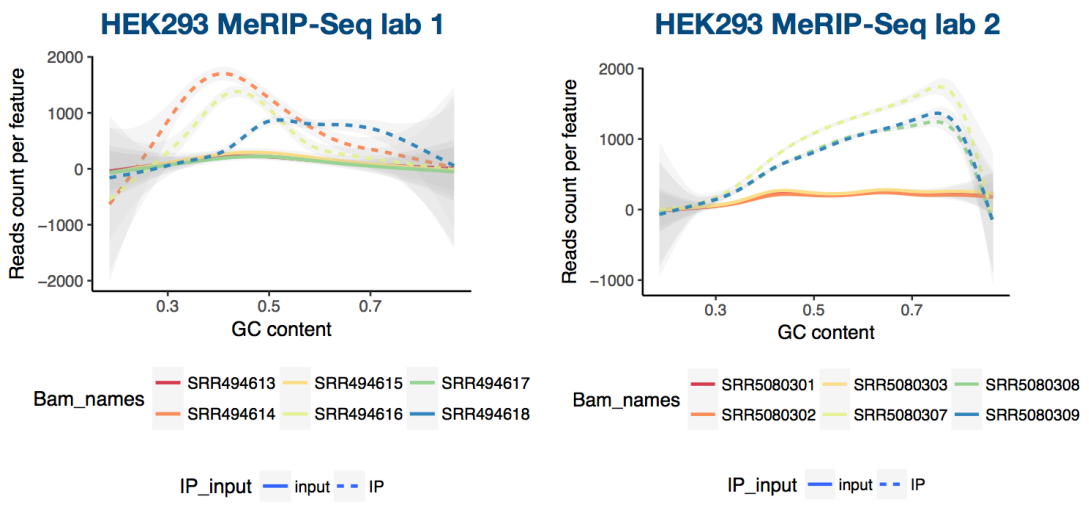


Figure2. The observed differences in GC content - reads count linear relationships between different laboratories under the same cell line condition. The curve on the graph represents the smooth regression (Gaussian Kernel Method) estimates of the linear relationship between GC content and the reads count, the features used for reads count are the single based m^6A sites collected from miCLIP/m6A-CLIP experiments. The IP samples are labelled as the dotted lines. We could observe that the IP sample has different linear relationships between different batches, while the GC content linear relationships are highly variable within the same experimental batch.

To reduce the variation introduced by the inconsistent peak lengths, the reads abundance is quantified using the experimentally verified single based m^6A sites which are the same set of sites of the positive training data in WHSTLE[106]. The linear relationships between reads abundance are different under different laboratory conditions for the same cell line condition of human liver cells (**Figure 2**). Such differences may be mainly attributed to the technical variation during the PCR amplification process of the RNA fragments under different laboratory conditions.

The different in the GC content linear relationships will directly lead to the deviance of the statistical inferences during the peak calling process (**Figure 3**), which may potentially lead to the false positive peak calls under regions with extreme levels of GC contents. Additionally, GC content bias will lead to systematic error during the process of modification level quantification, hence it will lead to biased result in differential modification analysis (**Figure 4**).

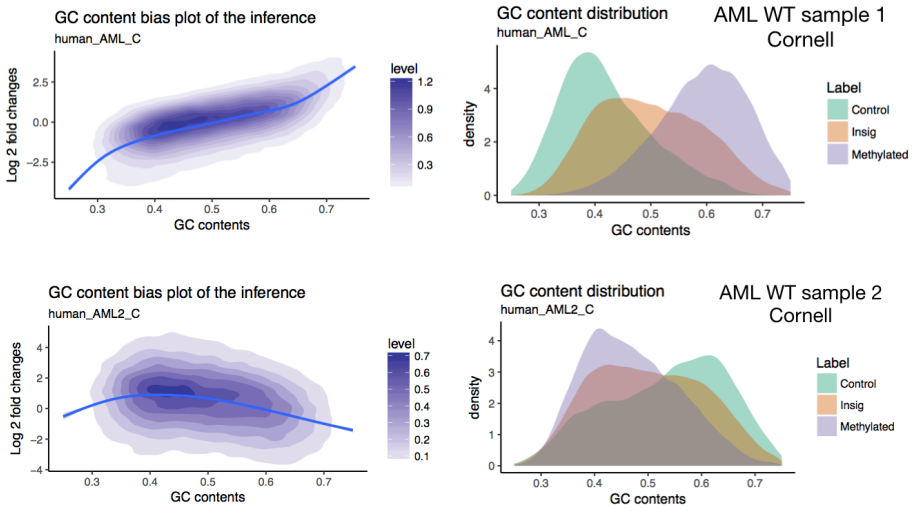


Figure3. The GC content biases observed in IP/input log2FC estimates and the inference of significant modified sites using DESeq2. The difference in GC content linear effects on the IP over input log2 Fold Change estimates is also observed, which suggests the DESeq2 model does not overcome the systematic biases introduced by the GC content. The inference over the significant modification also demonstrates the differences in the GC contents distributions among different directioanlities of the null hypothesis being rejected.

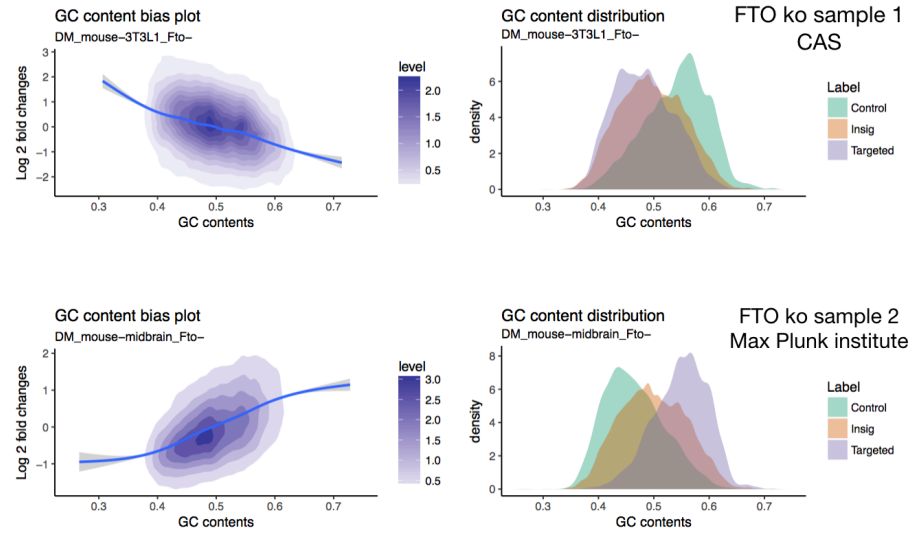


Figure4. The GC content biases observed in differential log2FC estimates and the inference over significantly differentially modified sites with DESeq2. As the quantification and inference of IP/input fold changes, the interactive effect estimates are strongly affected by the GC content. This phenomenon is

demonstrated in 2 datasets generated by different laboratories that are both investigating the effect of FTO knockdown on m^{6}A.

3.5. Peak Calling Performance Evaluation

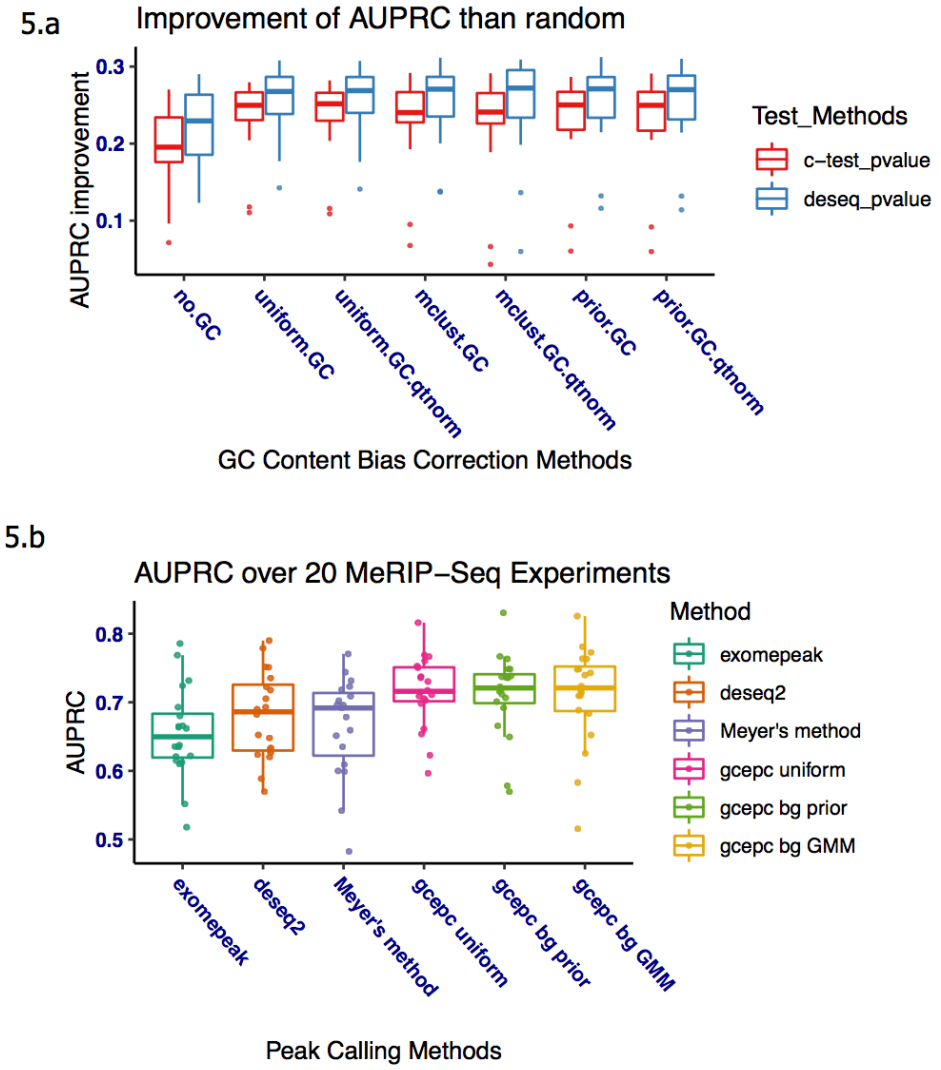


Figure5. Performance of peak calling evaluated on 20 m6A-Seq datasets. Using the site predicted by WHISTLE as the ground truth data, the performance of peak calling is evaluated by the AUPRC metrics. The Y axis in figure 5a and 5b indicates the improvement of AUROC compared with random predictions (random shuffle of the positive predicted bins). Figure 5a indicates that the GC content correction (labelled as GC on x-axis) can lead to the improvement of AUROC on both the c-test and DESeq2 test peak calling methods. However, the effects of quantile normalization (labelled as qtnorm on x-axis) and background GC effect estimation (labelled as bg prior & bg GMM) are weak in peak calling performance context. Figure 5b further confirmed the improvement of peak calling performances using Deseq2 + GC content normalization compared with exomePeak.

To evaluate the performance of peak calling on MeRIP-Seq datasets, we generated a set of predicted modification sites using WHISTLE classifier on mature RNA. The predictions of WHISTLE classifier are made on the whole exon region of the hg19 genome. All the RRACH sites with posterior probability > 0.5 are labeled as positive sites while the rest them are labeled as the negatives.

Using the default peak calling process of gcepc, exons are first divided into bins of 25bp in length. Different statistical tests are applied to infer the bins with significant modifications, and the p-value returned are filtered with different decision values to generate AUPRC (Area Under Precision-Recall Curve). When accounting for the overlap between the bins and m^6A sites, each bin is resized into the width of 201bp fixed by their centre. In our case, Recall is defined as the proportion of positive m6A sites overlapped with positive bins. Precision is defined as the proportion of positive bins overlapped with positive m6A sites.

After evaluating the prediction performance on 20 independent MeRIP-Seq experiment, the result showed that the methods implemented in gcepc have better prediction performances of m^6A site prediction compared with the traditional model used in exomePeak (**Figure 5**). Moreover, the DESeq2 method does have superior performances on peak calling compared with exomePeak over 20 samples. Importantly, the inclusion of GC content correction offset does improve the peak calling performance globally over the 20 samples. However, the estimation of GC content correction offsets on background will on average not lead to further performance improvement. This outcome might indicate that the biological selectivity of m^6A modification is not highly dependent on GC content.

3.6. Reduction of the batch effect for m^6A level quantification

The benefits of normalizing over GC content linear effects on reads abundance are not only to reduce the within group technical variance, it could also reduece the groupwise variance by eliminating the major source of technical variance correlated with

experiment batches. This phenomenon is demonstrated by a clustering analysis on 32 MeRIP-seq experiments from different laboratories (**Figure 6**). The modification levels of the 32 MeRIP-seq experiments are quantified using the $\log2(\frac{IP\ count+1}{input\ count+1})$ approach and the method developed in gcepc. The clustering is conducted after the methylation levels are rescaled by samples. The log2FC estimated using gcepc demonstrates greater within cluster similarities, also the clustering partition on gcepc group is more consistent with the real cell line label compared with the quantification made by the common method.

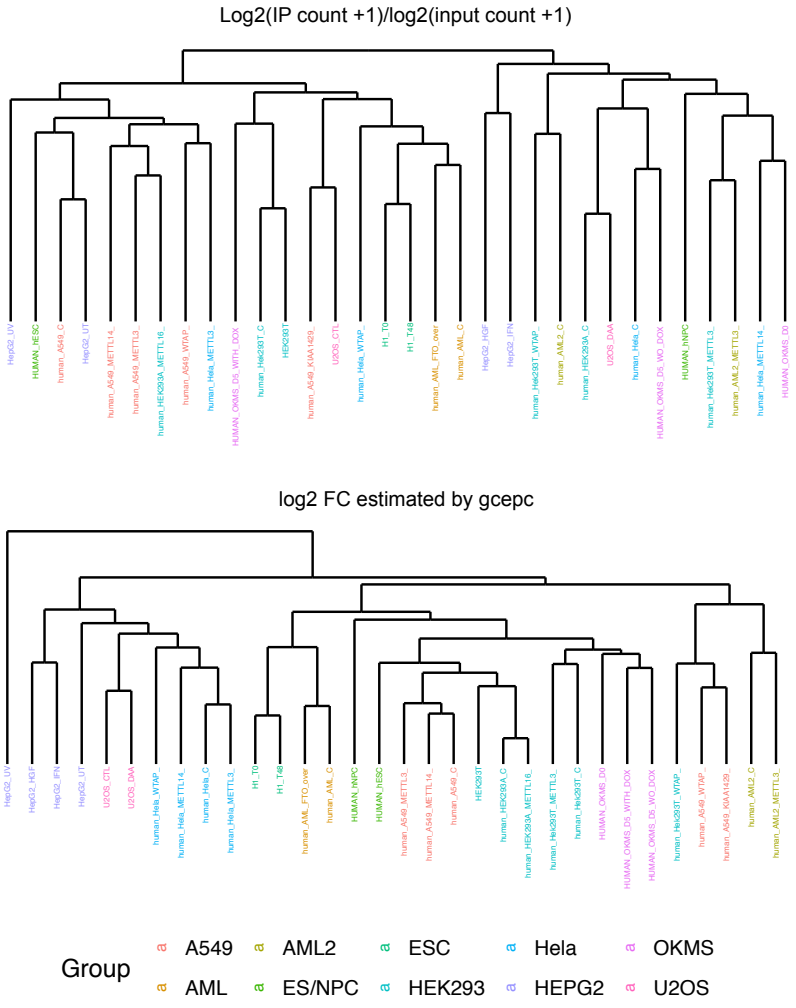


Figure6. Reduction of batch effect evaluated by the hierarchical clustering of samples. Compared with the traditional quantification method of log2 IP over input ratio. The dendrogram learned from the log2FC estimated by gcepc is the most consistent to the label of tissue information. Also, the within cluster distances for the latter is much smaller, which indicates the reduction of the within sample variations.

The significance of GC content correction can also be demonstrated by the clustering analysis of modification sites. Before clustering, the entries are first rescaled by columns and then rescaled by rows. The clustering analysis is conducted on the top 10000 modification sites having the highest variance. The clustering algorithm used is the K-medoids algorithm with Euclidean distance and K = 4. We applied the logistic regression

analysis using the 55 genomic features on each cluster separately. For the clustering result, the quantification derived by the common method results in the most explanatory features dominated by the GC content features, and the total deviances reduction for each model are significant smaller than the model fitted on the matrix quantified by gcepc. Clustering result for the second matrix is more biologically explainable. Almost all of the 4 clusters in the gcepc quantified analysis are associated with the RNA binding sites of the crucial m6A regulators.

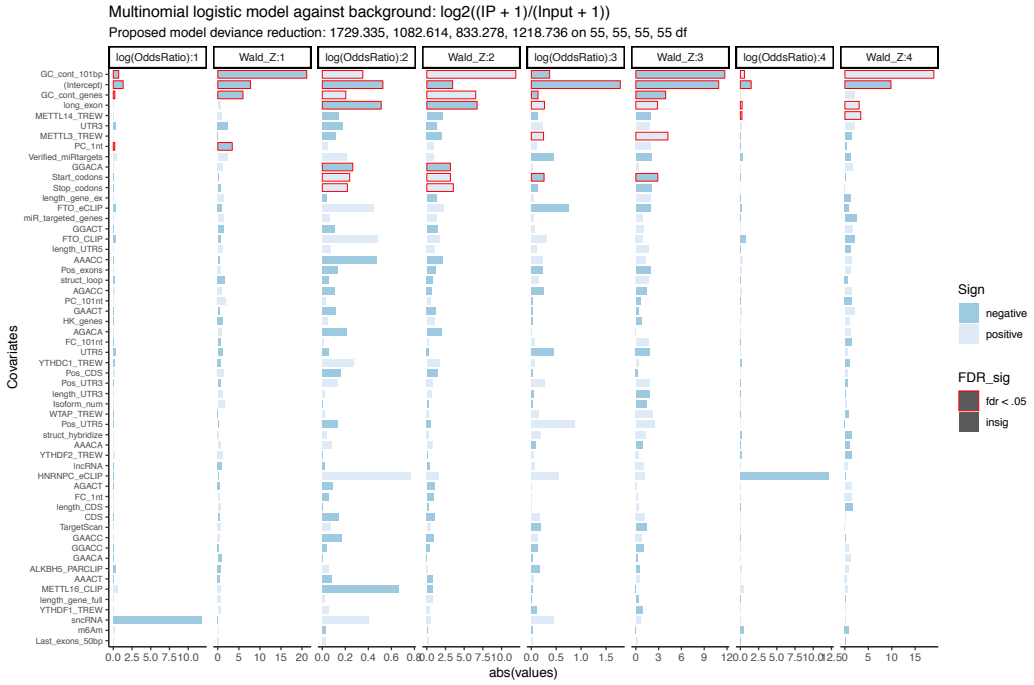


Figure7. Contribution of Genomic features in explaining the clustering results of modification sites. The modification levels quantified on single based modification are clustered by features (sites) with K = 4. The association between 54 genomic features and each cluster label is examined by a logistic regression model. The bar plot above indicates the logit (log odds) estimate for each genomic feature, the bars labelled by red colour are the significant association by the FDR adjusted 2 sided Wald test, and the colour of the bars indicate the direction of association. We could observe the clustering analysis using gcepc quantification yields much meaningful clusters compared with the analysis conducted on the quantities of log2((IP+1)/(input +1)).

In conclusion, the peak calling pipeline developed by gcepc could substantially improve the peak calling and quantification accuracy on MeRIP-Seq data. The improvement made is mostly due to the modeling over the biological variation and the GC content biases. In the future works, a more statistical rigorous method could be developed to unbiasedly estimate the technical GC content linear effect on IP samples, such as the one developed by gcapc[107] on CHIP-Seq data. Additionally, the estimation error of methylation level can be thoroughly studied with the different estimators (such as MLE and MAP of GLM) defined by DESeq2, a more rigorous bound could be used to decide the statistical reliability of the modification quantification on MeRIP-Seq data.

3.8 Materials and Data Availabilitie

ID	Sample Label	SRA Study	SRR RUN	Publication
1	HepG2-UV	SRP012098	SRR456542-SRR456543	p01 [77]
2	HepG2-HS		SRR456544-SRR456545	
3	HepG2-HGF		SRR456546-SRR456547	
4	HepG2-IFN		SRR456548-SRR456549	
5	HEK293T-1	SRP007335	SRR494613-SRR494618	p02 [102]
6	Hela	SRP022152	SRR847358-SRR847361, SRR847370-SRR847373	p03 [108]
7	Hela-METTL14-		SRR847362-SRR847365	
8	Hela-WTAP-		SRR847366-SRR847369	
9	Hela-METTL3-		SRR847374-SRR847377	
10	U2OS	SRP026127	SRR903368-SRR903370, SRR903374-SRR903376	p04 [109]
11	U2OS-DAA		SRR903371-SRR903373, SRR903377-SRR903379	
12	H1ESC	SRP033229	SRR1035213-SRR1035224	p05 [110]
13	H1ESC-T48		SRR1035217-SRR1035220	
14	hNPC	SRP039397	SRR1182582-SRR1182586	p06 [69]
15	hESC		SRR1182587-SRR1182590	
16	HEK293T-2-WTAP-		SRR1182591-SRR1182592	
17	HEK293T-2-METTL3-		SRR1182593-SRR1182594	
18	HEK293T-2		SRR1182595-SRR1182596	
19	OKMSfibro-Dox		SRR1182597-SRR1182598	
20	OKMSfibro		SRR1182599-SRR1182600	
21	OKMSiPC		SRR1182601-SRR1182602	
22	A549-WTAP-		SRR1182603-SRR1182606, SRR1182625-SRR1182626	
23	A549-METTL14-		SRR1182607-SRR1182614, SRR1182635-SRR1182636	
24	A549-METTL3-		SRR1182615-SRR1182618, SRR1182629-SRR1182630	

25	A549		SRR1182619-SRR1182624, SRR1182633-SRR1182634	
26	A549-KIAA1429-		SRR1182627-SRR1182628	
27	AML-1-FTO+	SRP067910	SRR3066062-SRR3066065	p07 [111]
28	AML-1		SRR3066066-SRR3066069	
29	gsc11		SRR4310464-SRR4310465, SRR4310468-SRR4310469	p08 [72]
30	gsc11-ALKBH5-		SRR4310466-SRR4310467, SRR4310470-SRR4310471	
31	HEK293A	SRP094637	SRR5080301-SRR5080303, SRR5080307-SRR5080309	p09 [112]
32	HEK293A-METTL16-		SRR5080304-SRR5080306, SRR5080310-SRR5080312	

Note: The data were downloaded directly from GEO.

Acknowledgement:

Thank for Dr. Subbarayalu Panneerdoss and Dr. Santosh Timilsina (University of Texas at San, San Antonio) and Dr. Jing-Ting Zhu (Xi'an Jiaotong-Liverpool University) for the experimental supports; Kun-Qi Chen for the build of prediction models in WHISTLE; Qing Zhang, Jia Lin Ma, and other bioinformatics students at XJTLU for the discussion of my studies; and Dr. Jia Meng and Dr. Rong Rong and Dr. Zhi Liang Lu for their supervision and support in my PhD research and thesis writing. The research presented in this thesis was supported by funds from National Natural Science Foundation of China [61401370 and 31671373 to JM, 81373469 to ZLL]; the Jiangsu Natural Science Foundation [BK20140403 to JM], the US National Institutes of Health [R01GM113245 to YH], IIMS Translational Technology Resource (TTR) Award to MKR and YC and NCI Cancer Center Shared Resources NCI P30CA54174 to YC. Part of BS-seq experiment was performed by the Genome Sequencing Facility of the Greehey Children's Cancer Research Institute, UTHSCSA.

Bibliography:

1. Suzuki MM, Bird A: **DNA methylation landscapes: provocative insights from epigenomics**. *Nat Rev Genet* 2008, **9**(6):465-476.
2. Robertson KD: **DNA methylation and human disease**. *Nat Rev Genet* 2005, **6**(8):597-610.
3. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, Soderhall C, Scheynius A, Kere J: **Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility**. *PLoS One* 2012, **7**(7):e41361.
4. Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B: **Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome**. *Cell* 2012, **148**(4):816-831.
5. He C: **Grand challenge commentary: RNA epigenetics?** *Nat Chem Biol* 2010, **6**(12):863-865.
6. Schibler U, Kelley DE, Perry RP: **Comparison of methylated sequences in messenger RNA and heterogeneous nuclear RNA from mouse L cells**. *Journal of molecular biology* 1977, **115**(4):695-714.
7. Desrosiers R, Friderici K, Rottman F: **Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells**. *Proc Natl Acad Sci U S A* 1974, **71**(10):3971-3975.
8. Dubin DT, Taylor RH: **The methylation state of poly A-containing messenger RNA from cultured hamster cells**. *Nucleic acids research* 1975, **2**(10):1653-1668.
9. Grosjean H: **Fine-tuning of RNA functions by modification and editing**: Springer; 2005.
10. Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C: **N 6-methyladenosine modulates messenger RNA translation efficiency**. *Cell* 2015, **161**(6):1388-1399.
11. Fustin JM, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I *et al*: **RNA-methylation-dependent RNA processing controls the speed of the circadian clock**. *Cell* 2013, **155**(4):793-806.
12. Alarcon CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF: **N6-methyladenosine marks primary microRNAs for processing**. *Nature* 2015, **519**(7544):482-485.
13. Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T: **N6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions**. *Nature* 2015, **518**(7540):560-564.
14. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G *et al*: **N6-methyladenosine-dependent regulation of messenger RNA stability**. *Nature* 2014, **505**(7481):117-120.
15. Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian S-B: **Dynamic m6A mRNA methylation directs translational control of heat shock response**. *Nature* 2015, **advance online publication**.

16. Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, Hershkovitz V, Peer E, Mor N, Manor YS: **m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation**. *Science* 2015:1261417.
17. Hussain S, Aleksic J, Blanco S, Dietmann S, Frye M: **Characterizing 5-methylcytosine in the mammalian epitranscriptome**. *Genome Biol* 2013, **14**(11):215.
18. Burgess AL, David R, Searle IR: **Conservation of tRNA and rRNA 5-methylcytosine in the kingdom Plantae**. *BMC Plant Biology* 2015, **15**(1):1-17.
19. Warren L, Manos PD, Ahfeldt T, Loh Y-H, Li H, Lau F, Ebina W, Mandal PK, Smith ZD, Meissner A: **Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA**. *Cell stem cell* 2010, **7**(5):618-630.
20. Zhang X, Liu Z, Yi J, Tang H, Xing J, Yu M, Tong T, Shang Y, Gorospe M, Wang W: **The tRNA methyltransferase NSun2 stabilizes p16INK4 mRNA by methylating the 3 [prime]-untranslated region of p16**. *Nature communications* 2012, **3**:712.
21. Chen Y, Sierzputowska-Gracz H, Guenther R, Everett K, Agris PF: **5-Methylcytidine is required for cooperative binding of magnesium (2+) and a conformational transition at the anticodon stem-loop of yeast phenylalanine tRNA**. *Biochemistry* 1993, **32**(38):10249-10253.
22. Motorin Y, Helm M: **tRNA stabilization by modified nucleotides**. *Biochemistry* 2010, **49**(24):4934-4944.
23. Tuorto F, Liebers R, Musch T, Schaefer M, Hofmann S, Kellner S, Frye M, Helm M, Stoecklin G, Lyko F: **RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis**. *Nature structural & molecular biology* 2012, **19**(9):900-905.
24. Schaefer M, Pollex T, Hanna K, Tuorto F, Meusburger M, Helm M, Lyko F: **RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage**. *Genes Dev* 2010, **24**.
25. Chan CT, Pang YLJ, Deng W, Babu IR, Dyavaiah M, Begley TJ, Dedon PC: **Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins**. *Nature communications* 2012, **3**:937.
26. Rai K, Chidester S, Zavala CV, Manos EJ, James SR, Karpf AR, Jones DA, Cairns BR: **Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish**. *Genes & development* 2007, **21**(3):261-266.
27. Wei C-M, Gershowitz A, Moss B: **Methylated nucleotides block 5′ terminus of HeLa cell messenger RNA**. *Cell* 1975, **4**(4):379-386.
28. Hussain S, Tuorto F, Menon S, Blanco S, Cox C, Flores JV, Watt S, Kudo NR, Lyko F, Frye M: **The mouse cytosine-5 RNA methyltransferase NSun2 is a component of the chromatoid body and required for testis differentiation**. *Molecular and cellular biology* 2013, **33**(8):1561-1570.
29. Blanco S, Dietmann S, Flores JV, Hussain S, Kutter C, Humphreys P, Lukk M, Lombard P, Treps L, Popis M: **Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders**. *The EMBO journal* 2014, **33**(18):2020-2039.

30. Chow CS, Lamichhane TN, Mahto SK: **Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications**. *ACS chemical biology* 2007, **2**(9):610-619.
31. Schaefer M: **Chapter Fourteen - RNA 5-Methylcytosine Analysis by Bisulfite Sequencing**. In: *Methods in Enzymology.* Edited by Chuan H, vol. Volume 560: Academic Press; 2015: 297-329.
32. Schaefer M, Pollex T, Hanna K, Lyko F: **RNA cytosine methylation analysis by bisulfite sequencing**. *Nucleic Acids Res* 2009, **37**(2):e12.
33. Motorin Y, Lyko F, Helm M: **5-methylcytosine in RNA: detection, enzymatic formation and biological functions**. *Nucleic Acids Res* 2010, **38**(5):1415-1430.
34. Khoddami V, Cairns BR: **Identification of direct targets and modified bases of RNA cytosine methyltransferases**. *Nature biotechnology* 2013, **31**(5):458-464.
35. Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, Parker BJ, Suter CM, Preiss T: **Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA**. *Nucleic acids research* 2012:gks144.
36. Amort T, Rieder D, Wille A, Khokhlova-Cubberley D, Riml C, Trixl L, Jia X-Y, Micura R, Lusser A: **Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain**. *Genome Biology* 2017, **18**(1):1.
37. Yang X, Yang Y, Sun B-F, Chen Y-S, Xu J-W, Lai W-Y, Li A, Wang X, Bhattarai DP, Xiao W *et al*: **5-methylcytosine promotes mRNA export [mdash] NSUN2 as the methyltransferase and ALYREF as an m5C reader**. *Cell Res* 2017, **27**(5):606-625.
38. Adams JM, Cory S: **Modified nucleosides and bizarre 5'-termini in mouse myeloma mRNA**. *Nature* 1975, **255**(5503):28-33.
39. Salditt-Georgieff M, Jelinek W, Darnell JE, Furuichi Y, Morgan M, Shatkin A: **Methyl labeling of HeLa cell hnRNA: a comparison with mRNA**. *Cell* 1976, **7**(2):227-237.
40. Lee E-J, Pei L, Srivastava G, Joshi T, Kushwaha G, Choi J-H, Robertson KD, Wang X, Colbourne JK, Zhang L *et al*: **Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing**. *Nucleic Acids Research* 2011, **39**(19):e127.
41. Amort T, Rieder D, Wille A, Khokhlova-Cubberley D, Riml C, Trixl L, Jia XY, Micura R, Lusser A: **Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain**. *Genome Biology* 2017, **18**(1):1.
42. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, Maldotti M, Anselmi F, Oliviero S: **Intragenic DNA methylation prevents spurious transcription initiation**. *Nature* 2017, **543**(7643):72.
43. Krueger F: **Trim Galore**. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* 2015.
44. Rieder D, Amort T, Kugler E, Lusser A, Trajanoski Z: **meRanTK: methylated RNA analysis ToolKit**. *Bioinformatics* 2016, **32**(5):782-785.
45. Hofacker IL: **Vienna RNA secondary structure server**. *Nucleic acids research* 2003, **31**(13):3429-3431.
46. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for Computing and Annotating Genomic Ranges**. *PLoS Comput Biol* 2013, **9**(8):e1003118.
47. Cui X, Wei Z, Zhang L, Liu H, Sun L, Zhang S-W, Huang Y, Meng J: **Guitar: An R/Bioconductor Package for Gene Annotation Guided Transcriptomic**

Analysis of RNA-Related Genomic Features. *BioMed Research International* 2016, **2016**:8.
48. Wickham H: **ggplot2: elegant graphics for data analysis**: Springer Science & Business Media; 2009.
49. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM: **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis**. *BMC Bioinformatics* 2010, **11**(1):1-9.
50. Liu L: **QNB: Differential RNA Methylation Analysis for Count-Based Small-Sample Sequencing Data with a Quad-Negative Binomial Model**. *CRAN R package* 2017.
51. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biology* 2014, **15**(12):550.
52. Zao CL, Ward JA, Tomanek L, Cooke A, Berger R, Armstrong K: **Virological and serological characterization of SRV-4 infection in cynomolgus macaques**. *Archives of virology* 2011, **156**(11):2053-2056.
53. Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE, Bestor TH: **Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2**. *Science* 2006, **311**(5759):395-398.
54. Qiu P, Zhang L: **Identification of markers associated with global changes in DNA methylation regulation in cancers**. *BMC bioinformatics* 2012, **13**(Suppl 13):S7.
55. Camara Y, Asin-Cayuela J, Park CB, Metodiev MB, Shi YH, Ruzzenente B, Kukat C, Habermann B, Wibom R, Hultenby K *et al*: **MTERF4 Regulates Translation by Targeting the Methyltransferase NSUN4 to the Mammalian Mitochondrial Ribosome**. *Cell Metabolism* 2011, **13**(5):527-539.
56. Pardal R, Clarke MF, Morrison SJ: **Applying the principles of stem-cell biology to cancer**. *Nature Reviews Cancer* 2003, **3**(12):895-902.
57. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control**. *Nature* 2011, **473**(7347):337-342.
58. Wang X, Gu J, Hilakivi-Clarke L, Clarke R, Xuan J: **DM-BLD: Differential methylation detection using a hierarchical Bayesian model exploiting local dependency**. *Bioinformatics* 2016:btw596.
59. Klein H-U, Hebestreit K: **An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data**. *Briefings in bioinformatics* 2015:bbv095.
60. Li S, Garrett-Bakelman FE, Akalin A, Zumbo P, Levine R, To BL, Lewis ID, Brown AL, D'Andrea RJ, Melnick A: **An optimized algorithm for detecting and annotating regional differential methylation**. *BMC bioinformatics* 2013, **14**(Suppl 5):S10.
61. Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, Zhao BS, Mesquita A, Liu C, Yuan CL *et al*: **Recognition of RNA N6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation**. *Nature Cell Biology* 2018, **20**(3):285-295.
62. Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T: **N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions**. *Nature* 2015, **518**(7540):560-564.
63. Legrand C, Tuorto F, Hartmann M, Liebers R, Jacob D, Helm M, Lyko F: **Statistically robust methylation calling for whole-transcriptome bisulfite**

sequencing reveals distinct methylation patterns for mouse RNAs. *Genome Research* 2017, **27**(9).

64. Li X, Xiong X, Wang K, Wang L, Shu X, Ma S, Yi C: **Transcriptome-wide mapping reveals reversible and dynamic N1-methyladenosine methylome**. *Nature Chemical Biology* 2016, **12**(5).
65. Zhao BS, Roundtree IA, He C: **Post-transcriptional gene regulation by mRNA modifications**. *Nat Rev Mol Cell Biol* 2017, **18**.
66. Liu H, Flores MA, Meng J, Zhang L, Zhao X, Rao MK, Chen Y, Huang Y: **MeT-DB: a database of transcriptome methylation in mammalian cells**. *Nucleic Acids Research* 2015, **43**(Database issue):D197.
67. Sun WJ, Li JH, Liu S, Wu J, Zhou H, Qu LH, Yang JH: **RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data**. *Nucleic Acids Research* 2015, **44**(D1):D259.
68. Fu Y, Dominissini D, Rechavi G, He C: **Gene expression regulation mediated through reversible m(6)a RNA methylation**. *Nat Rev Genet* 2014, **15**.
69. Schwartz S, Mumbach M, Jovanovic M, Wang T, Maciag K, Bushkin GG, Mertins P, Ter-Ovanesyan D, Habib N, Cacchiarelli D: **Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites**. *Cell Reports* 2014, **8**(1):284-296.
70. Xiang W, Jing F, Yuan X, Guan Z, Zhang D, Zhu L, Zhou G, Qiang W, Huang J, Tang C: **Structural basis of N6-adenosine methylation by the METTL3–METTL14 complex**. *Nature* 2016, **534**(7608).
71. Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, Yi C, Lindahl T, Pan T, Yang YG: **N6-Methyladenosine in Nuclear RNA is a Major Substrate of the Obesity-Associated FTO**. *Nature Chemical Biology* 2011, **7**(12):885-887.
72. Zhang S, Zhao BS, Zhou A, Lin K, Zheng S, Lu Z, Chen Y, Sulman EP, Xie K, Bögler O: **m 6 A demethylase ALKBH5 maintains tumorigenicity of glioblastoma stem-like cells by sustaining FOXM1 expression and cell proliferation program**. *Cancer cell* 2017, **31**(4):591-606. e596.
73. Li F, Zhao D, Wu J, Shi Y: **Structure of the YTH domain of human YTHDF2 in complex with an m6A mononucleotide reveals an aromatic cage for m6A recognition**. *Cell Research* 2014, **24**(12):1490-1492.
74. Zou S, Toh JD, Wong KH, Gao YG, Hong W, Woon EC: **N(6)-Methyladenosine: a conformational marker that regulates the substrate specificity of human demethylases FTO and ALKBH5**. *Scientific Reports* 2016, **6**:25677.
75. **Machnicka MA, Milanowska K, Oglou OO, Purta E, Kurkowska M, Olchowik A, Januszewski W, Kalinowski S, Dunin-Horkawicz S, Rother KM: MODOMICS: a database of RNA modification pathways—2012 update. Nucleic acids research 2012:gks1007**. In.
76. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser**. *Genome research* 2009, **19**(9):1630-1638.
77. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M: **Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq**. *Nature* 2012, **485**(7397):201-206.
78. Meyer KD, Yogesh S, Paul Z, Olivier E, Mason CE, Jaffrey SR: **Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons**. *Cell* 2012, **149**(7):1635–1646.

79. Meng J, Lu Z, Liu H, Zhang L, Zhang S, Chen Y, Rao MK, Huang Y: **A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package**. *Methods* 2014, **69**(3):274-281.
80. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**.
81. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods* 2015, **12**.
82. Zhao X, Yang Y, Sun BF, Shi Y, Yang X, Xiao W, Hao YJ, Ping XL, Chen YS, Wang WJ: **FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis**. *Cell Research* 2014, **24**(12):1403-1419.
83. Siepel A, Haussler D: **Phylogenetic hidden Markov models**. In: *Statistical methods in molecular evolution.* Springer; 2005: 325-351.
84. Gulko B, Gronau I, Hubisz MJ, Siepel A: **Probabilities of Fitness Consequences for Point Mutations Across the Human Genome**. *bioRxiv* 2014:006825.
85. Gruber AR, Bernhart SH, Lorenz R: **The ViennaRNA web services**. In: *RNA bioinformatics.* Springer; 2015: 307-326.
86. Betel D, Koppal A, Agius P, Sander C, Leslie C: **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. *Genome biology* 2010, **11**(8):R90.
87. Agarwal V, Bell GW, Nam JW, Bartel DP: **Predicting effective microRNA target sites in mammalian mRNAs**. *eLife* 2015, **4**.
88. Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vågbø CB, Geula S, Hanna JH, Black DL, Darnell JE, Darnell RB: **m6A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover**. *Genes & development* 2017, **31**(10):990-1006.
89. Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL: **ViennaRNA Package 2.0**. *Algorithms for Molecular Biology* 2011, **6**(1):26.
90. Eisenberg E, Levanon EY: **Human housekeeping genes, revisited**. *Trends in Genetics* 2013, **29**(10):569-574.
91. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H: **miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions**. *Nucleic acids research* 2017, **46**(D1):D296-D302.
92. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57.
93. Betel D, Koppal A, Agius P, Sander C, Leslie C: **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. *Genome biology* 2010, **11**(8):R90.
94. Agarwal V, Bell GW, Nam J-W, Bartel DP: **Predicting effective microRNA target sites in mammalian mRNAs**. *elife* 2015, **4**:e05005.
95. Liu H, Yue D, Chen Y, Gao SJ, Huang Y: **Improving performance of mammalian microRNA target prediction**. *BMC bioinformatics* 2010, **11**:476.
96. Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko MT, Hwang JK: **KinasePhos 2.0: a web server for identifying protein**

kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic acids research* 2007, **35**(Web Server issue):W588-594.
97. Chen W, Tang H, Lin H: **MethyRNA: a web server for identification of N(6)-methyladenosine sites**. *Journal of biomolecular structure & dynamics* 2017, **35**(3):683-687.
98. Xiang S, Liu K, Yan Z, Zhang Y, Sun Z: **RNAMethPre: A Web Server for the Prediction and Query of mRNA m6A Sites**. *PLoS One* 2016, **11**(10):e0162707.
99. Chang C-C, Lin C-J: **LIBSVM: A library for support vector machines**. *ACM Trans Intell Syst Technol* 2011, **2**(3):1-27.
100. Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE: **The birth of the Epitranscriptome: deciphering the function of RNA modifications**. *Genome Biology* 2012, **13**(10):1-12.
101. Dominissini D, Moshitch-Moshkovitz S, Salmon-Divon M, Amariglio N, Rechavi G: **Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing**. *Nat Protoc* 2013, **8**.
102. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR: **Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons**. *Cell* 2012, **149**.
103. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol* 2008, **9**.
104. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biology* 2010, **11**(10):R106.
105. Hansen KD, Irizarry RA, Wu Z: **Removing technical variability in RNA-seq data using conditional quantile normalization**. *Biostatistics* 2012, **13**(2):204-216.
106. Chen K, Wei Z, Zhang Q, Wu X, Rong R, Lu Z, Su J, de Magalhães JP, Rigden DJ, Meng J: **WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach**. *Nucleic Acids Research* 2019.
107. Teng M, Irizarry RA: **Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data**. *Genome Research* 2017, **27**(11):1930.
108. Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X: **A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation**. *Nature chemical biology* 2014, **10**(2):93-95.
109. Fustin JM, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I: **RNA-methylation-dependent RNA processing controls the speed of the circadian clock**. *Cell* 2013, **155**.
110. Batista PJ, Molinie B, Wang J, Qu K, Zhang J, Li L, Bouley DM, Lujan E, Haddad B, Daneshvar K: **m 6 A RNA modification controls cell fate transition in mammalian embryonic stem cells**. *Cell stem cell* 2014, **15**(6):707-719.
111. Li Z, Weng H, Su R, Weng X, Zuo Z, Li C, Huang H, Nachtergaele S, Dong L, Hu C: **FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N 6 -Methyladenosine RNA Demethylase**. *Cancer Cell* 2016.

112. Pendleton KE, Chen B, Liu K, Hunter OV, Xie Y, Tu BP, Conrad NK: **The U6 snRNA m 6 A methyltransferase METTL16 regulates SAM synthetase intron retention**. *Cell* 2017, **169**(5):824-835. e814.
113. Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.